Resource Allocation in Stochastic Processing Networks: Performance and Scaling

by

Yuan Zhong

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2012

Author .....
Sloan School of Management
August 17, 2012

Certified by .....
Devavrat Shah
Jamieson Associate Professor,
Department of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by .....
John N. Tsitsiklis
Clarence J. Lebel Professor of Electrical Engineering,
Department of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Dimitris Bertsimas
Boeing Leaders for Global Operations Professor,
Sloan School of Management,
Co-director, Operations Research Center

Resource Allocation in Stochastic Processing Networks: Performance and Scaling

by

Yuan Zhong

Submitted to the Sloan School of Management
on August 17, 2012, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

This thesis addresses the design and analysis of resource allocation policies in large-scale stochastic systems, motivated by examples such as the Internet, cloud facilities, wireless networks, etc. A canonical framework for modeling many such systems is provided by "stochastic processing networks" (SPN) (Harrison [28, 29]). In this context, the key operational challenge is efficient and timely resource allocation.

We consider two important classes of SPNs: switched networks and bandwidth-sharing networks. Switched networks are constrained queueing models that have been used successfully to describe the detailed packet-level dynamics in systems such as input-queued switches and wireless networks. Bandwidth-sharing networks have primarily been used to capture the long-term behavior of the flow-level dynamics in the Internet. In this thesis, we develop novel methods to analyze the performance of existing resource allocation policies, and we design new policies that achieve provably good performance.

First, we study performance properties of so-called Maximum-Weight-α (MW-α) policies in switched networks, and of α-fair policies in bandwidth-sharing networks, both of which are well-known families of resource allocation policies, parametrized by a positive parameter $\alpha > 0$. We study both their transient properties as well as their steady-state behavior.

In switched networks, under a MW-α policy with $\alpha \geq 1$, we obtain bounds on the maximum queue size over a given time horizon, by means of a maximal inequality derived from the standard Lyapunov drift condition. As a corollary, we establish the full state space collapse property when $\alpha \geq 1$. In the steady-state regime, for any $\alpha > 0$, we obtain explicit exponential tail bounds on the queue sizes, by relying on a norm-like Lyapunov function, different from the standard Lyapunov function used in the literature.

Methods and results are largely parallel for bandwidth-sharing networks. Under an α-fair policy with $\alpha \geq 1$, we obtain bounds on the maximum number of flows in the network over a given time horizon, and hence establish the full state space collapse property when $\alpha \geq 1$. In the steady-state regime, using again a norm-like

Lyapunov function, we obtain explicit exponential tail bounds on the number of flows, for any $\alpha > 0$. As a corollary, we establish the validity of the diffusion approximation developed by Kang et al. [32], in steady state, for the case $\alpha = 1$.

Second, we consider the design of resource allocation policies in switched networks. At a high level, the central performance questions of interest are: what is the optimal scaling behavior of policies in large-scale systems, and how can we achieve it?

More specifically, in the context of general switched networks, we provide a new class of online policies, inspired by the classical insensitivity theory for product-form queueing networks, which admits explicit performance bounds. These policies achieve optimal queue-size scaling, in the conventional heavy-traffic regime, for a class of switched networks, thus settling a conjecture (documented in [51]) on queue-size scaling in input-queued switches.

In the particular context of input-queued switches, we consider the scaling behavior of queue sizes, as a function of the port number n and the load factor ρ. In particular, we consider the special case of uniform arrival rates, and we focus on the regime where $\rho = 1 - 1/f(n)$, with $f(n) \geq n$. We provide a new class of policies under which the long-run average total queue size scales as $O(n^{1.5} f(n) \log f(n))$. As a corollary, when $f(n) = n$, the long-run average total queue size scales as $O(n^{2.5} \log n)$. This is a substantial improvement upon prior works [44], [52], [48], [39], where the same quantity scales as $O(n^3)$ (ignoring logarithmic dependence on n).

Thesis Supervisor: Devavrat Shah
Title: Jamieson Associate Professor,
Department of Electrical Engineering and Computer Science

Thesis Supervisor: John N. Tsitsiklis
Title: Clarence J. Lebel Professor of Electrical Engineering,
Department of Electrical Engineering and Computer Science

Acknowledgments[1]

No words can fully express my gratitude to my advisors, Devavrat Shah and John Tsitsiklis, for their guidance and support throughout the course of my doctoral study at MIT. They have taught me so many things, and have shaped me both as a person and as a researcher. I have learned from them (and am still learning!) basic principles of good research (such as the importance of simplicity and clarity), which often extend to wisdom of life. They have always demonstrated the utmost passion for life and research, and the highest level of scientific rigor. Indeed, it would be hard for me to imagine where I would stand today without them, let alone the possibility of this thesis.

I would like to thank David Gamarnik and Kavita Ramanan, for taking the time to serve on my committee, and for their constructive feedback on the thesis. I am also grateful to David for various technical interactions that have fostered my intellectual growth at MIT.

Chapter 5 of this thesis is based on a joint paper with Neil Walton, and I would like to thank him for a pleasurable collaboration and fulfilling learning experience.

The ORC and LIDS community have always been supportive. I would especially like to thank Vivek Farias for his help and support, his mentorship on teaching, and many thought-provoking conversations. I would also like to thank the staff members, Andrew Carvalho, Laura Rose, and Lynne Dell, for being helpful on numerous occasions.

Finally, I would like to dedicate this thesis to my parents, Nanyan Yu and Yongle Zhong. They have always been there to love and support me, and to them I express my deepest gratitude.

[1]I gratefully acknowledge the financial support of NSF Theoretical Foundation Collaborative Project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context

Modern processing systems, such as the Internet, cloud facilities, call centers, global supply chains, etc, are becoming increasingly complex. For example, a cloud facility could consist of tens of thousands of interconnected processors, and a call center of thousands of staff with overlapping skill sets. Common to these large-scale processing facilities is an enormous number of processing activities, coupled by resource constraints, and the key operational challenge faced by a system manager is efficient and timely resource allocation. Given the scale of such systems, no longer can a modern-day system manager depend purely on heuristics or experience to guide day-to-day operations; a rigorous and scientific approach is called for.

A canonical framework that has emerged over the past decade, for modeling many of the systems mentioned above, is provided by "stochastic processing networks" (SPN) (Harrison [28, 29]). SPNs capture details of a broad spectrum of networked systems at a fine granularity. Some examples of SPNs include:

(a) multi-class queueing networks, which have been used to model, for example, the process of wafer fabrication in manufacturing systems;

(b) parallel-server systems, which have been used to model, for example, operations of call centers;

(c) bandwidth-sharing networks, which have been used to model, for example, the flow-level dynamics in the Internet; and

(d) switched networks, which have been used to model, for example, the switching architecture in Internet routers, and wireless medium access.

In this thesis, we focus on switched networks and bandwidth-sharing networks, and we address the question of efficient resource allocation in these contexts.

An *ideal* resource allocation policy ought to be practically implementable and guarantee good practical and theoretical performance. Implementability means that the policy has low complexity, and good theoretical performance means that the policy has provably optimal bounds with respect to various performance metrics of interest. Because of the complexity of generic SPNs, the design of ideal policies remains an important challenge. In this thesis, we focus on performance-related questions, and believe that progress made in this thesis is a substantial advancement toward meeting this challenge.

More specifically, we are interested in advancing existing methods of performance analysis, and hence generating new insights into existing resource allocation policies considered in the literature, as well as designing new policies to achieve good performance bounds. For both of these objectives, we will see that system performance depends crucially on both the network structure and the system load. Indeed, this is a prominent theme of the thesis. For example, in scaling analysis of switched networks, the key question of interest is the optimal performance behavior of policies in large-scale networks, and we consider scaling behavior with respect to *both* the network structure and the system load. This is in contrast with previous works, which only consider the scaling behavior with respect to load.

In the next section, we review relevant research on the design and analysis of resource allocation policies, in the context of both switched networks and bandwidth-sharing networks, and highlight some open questions that need to be addressed.

1.2 Literature Review

Here we give an overview on the existing performance analysis methods and results for popular resource allocation policies in switched networks and in bandwidth-sharing networks. We first describe some of the policies proposed in the literature, which are most relevant to this thesis, including Maximum-Weight-α policies and α-fair bandwidth-sharing policies, and then give an account of their various known performance properties.

Existing Policies. In switched networks, of particular interest is the class of Maximum-Weight-α (MW-α) policies, parametrized by $\alpha > 0$. MW-α is the only known class of simple policies with universal performance guarantees. They were first introduced in the context of ad hoc wireless networks by Tassiulas and Ephremides [61], and in the context of input-queued switches by McKeown et al. [38, 41]. In a general version of the policy, a service vector with the maximum weight is chosen at each time step, with the weight of the service vector being equal to the sum of weights of the queues corresponding to this service. In particular, for each $\alpha > 0$, and for the choice of weight function $f_\alpha : x \mapsto x^\alpha$ the resulting policy is called the MW-α policy. The MW-α policies have served as an important guide for designing implementable policies for input-queued switches and wireless medium access [24, 25, 40, 46, 60]. Due to their importance, there has been a large body of research on their performance properties, which will be detailed shortly.

In bandwidth-sharing networks, α-fair policies have been studied extensively. At each time instant, the optimal solution to a concave utility maximization problem is chosen as the service rate vector, with the class of utility functions being parametrized by a positive parameter α. The general framework of utility maximization was proposed by Kelly, Mauloo and Tan [37] for a static version of the bandwidth-sharing problem, partly to understand the behavior of TCP in the Internet. Mo and Walrand [43] then introduced the class of "fair" bandwidth-sharing policies, parametrized by $\alpha > 0$. Subsequently, this theory has in turn motivated the invention of specialized congestion control mechanisms, such as FAST TCP [31]. The α-fair policies have also

generated a large body of research, to be detailed shortly.

Throughput Analysis. In both types of networks, the most basic performance question concerns necessary and sufficient conditions for stability, that is, for the existence of a steady-state distribution for the associated Markov chain (process). In switched networks, for MW-α policies, under fairly general assumptions on stochastic primitives, stability has been established for any $\alpha > 0$ [1, 13, 41, 61]. In bandwidth-sharing networks, stability was established by Bonald and Massoulié [6] for the case of α-fair policies with $\alpha > 0$, and by de Veciana et al. [15] for the case of max-min fairness ($\alpha \to \infty$) and proportionally fair policies ($\alpha = 1$). In all these scenarios, the stability conditions turned out to be the natural deterministic conditions based on mean arrival rates and service rates.

Bounds on Steady-State Queue Size. Given these stability results, the natural next question is whether the steady-state expectation of the total queue size in switched networks, or that of the total number of flows in bandwidth-sharing networks, is finite, and if so, to identify some non-trivial upper bounds. When $\alpha \geq 1$, the finiteness question can be answered in the affirmative, for both MW-α and α-fair policies, and explicit bounds can be obtained, by exploiting the same Lyapunov drift inequalities that had been used in earlier work to establish stability. However, this approach does not seem to apply to the case where $\alpha \in (0, 1)$, which remained an open problem; this is one of the problems that we settle in this thesis.

We remark that in switched networks, for MW-α policies with $\alpha \geq 1$, the explicit upper bounds take the form $c\frac{N}{1-\rho}$ [41, 51, 61], where c is a universal constant, N is the number of queues in the system, and ρ is the *load*. In many switched networks, such as input-queued switches, one can obtain universal lower bounds on the steady-state expectation of the total queue size, that have a much better dependence on N (for example, the lower bounds in input-queued switches take the form $\frac{\sqrt{N}}{1-\rho}$). In the particular instance of input-queued switches, a batching policy was proposed in [44] which gave an upper bound of $O(\sqrt{N}\log N/(1-\rho)^2)$. It is an open problem whether

this gap between the upper and lower bounds can be closed, and indeed, there are several forms of conjectures associated with this problem [51]. In this thesis, we settle one of the conjectures documented in [51].

Large-Deviations Analysis. Given that the Markov chain/process describing the underlying system has a steady-state distribution, another question concerns exponentially decaying bounds on the tail of the distribution. For both MW-α policies and α-fair policies, we provide results of this form, together with explicit bounds for the associated exponent. Related recent results include works by Stolyar [57] and by Venkataramanan and Lin [62], who provide a precise asymptotic characterization of the exponent of the tail probability, in steady state, for the case of switched networks (as opposed to flow-level network models). (To be precise, their results concern the $(1+\alpha)$ norm of the vector of queue sizes under maximum weight or pressure policies parametrized by $\alpha > 0$.) While exact, their results involve a variational characterization that appears to be difficult to evaluate (or even bound) explicitly. We also take note of work by Subramanian [59], who establishes a large deviations principle for a class of switched network models under maximum weight or pressure policies with $\alpha = 1$.

Let us provide here some remarks on the optimal tail exponent. For concreteness, we restrict our discussion to input-queued switches. In an input-queued switch, a universal lower bound of $-2(1-\rho)/\rho$ on the tail exponent of the steady-state queue-size distribution can be readily established, where ρ is the effective load on the system. In contrast, for MW-α policies, the explicit upper bounds that we obtain for the tail exponent have an additional dependence on N, the number of queues, and it is of interest to see whether MW-α policies, or for that matter, any policy, can close this gap between the upper and lower bounds. We settle this problem in this thesis, for a class of switched networks including input-queued switches.

Heavy-Traffic Analysis. The analysis of the steady-state distribution for underloaded networks provides only partial insights about the transient behavior of the

associated Markov chain/process. As a refinement, an important performance analysis method that has emerged over the past few decades focuses on the heavy-traffic regime, in which the system is critically loaded. The heavy-traffic (or diffusion) scaling of the network can lead to parsimonious approximations for the transient behavior. A general two-stage program for developing such diffusion approximations has been put forth by Bramson [9] and Williams [64], and has been carried out in detail for certain classes of queueing network models. To carry out this program, one needs to: (i) carry out a detailed analysis of a related fluid model when the network is critically loaded; and, (ii) identify a unique distributional limit of the associated diffusion-scaled processes by studying a related Skorohod problem.

For a general switched network, in [58], a complete characterization of the diffusion approximation for the queue-size process has been obtained, under a condition known as *"complete resource pooling,"* when the network is operating under the MW-α policy, for any $\alpha > 0$. The diffusion approximation was shown to be a one-dimensional diffusion process, where all components are equal at all time instance. Furthermore, it was established in [58] that under the heavy-traffic scaling (with complete resource pooling), a MW-α policy, for any $\alpha > 0$, minimizes the rescaled workload induced by any policy which establishes a form of heavy-traffic optimality of MW-α. However, complete resource pooling effectively requires that the underlying network have only one bottleneck, and does not capture the effect of the network structure.

In order to capture the dependence of queue sizes on the network structure, a heavy-traffic analysis of switched networks with multiple bottlenecks (without resource pooling) was pursued by Shah and Wischik [55]. They first carried out the first stage of the Bramson-Williams program outlined above, identifying the invariant manifolds of the associated critically loaded fluid models, for all $\alpha > 0$. They established the so-called multiplicative state space collapse, and identified a member, denoted by MW-0^+ (obtained by taking $\alpha \to 0$), of the class of maximum-weight policies as optimal with respect to a critical fluid model. State space collapse roughly means that in the heavy-traffic limit, and under diffusion scaling, the system state evolves in a much lower-dimensional space. Building upon the work of Shah and Wis-

chik [55], Kang and Williams [33] recently established a full diffusion approximation for the MW policy ($\alpha = 1$), in the particular case of input-queued switches.

For bandwidth-sharing networks, the first stage of the Bramson-Williams program was carried out by Kelly and Williams [36] who identified the invariant manifold of the associated critically loaded fluid model. This further led to the proof by Kang et al. [32] of a multiplicative state space collapse property, similar to results by Bramson [9]. We note that the above summarized results hold under α-fair policies with an arbitrary $\alpha > 0$. The second stage of the program has been carried out for the unweighted proportionally fair policy ($\alpha = 1$) by Kang et al. [32], under a technical *local traffic* condition (different from complete resource pooling, not necessarily weaker), and more recently, by Ye and Yao [65], under a somewhat less restrictive technical condition.

We note that when $\alpha \neq 1$, a diffusion approximation has not been established for either switched networks or bandwidth-sharing networks. In this case, it is of interest to see at least whether properties that are stronger than multiplicative state space collapse can be derived, something that is accomplished in this thesis.

The above outlined diffusion approximation results involve rigorous statements on the finite-time behavior of the original process. Kang et al. [32] further established that for the particular setting that they consider, the resulting diffusion approximation has an elegant product-form steady-state distribution; this result gives rise to an intuitively appealing interpretation of the relation between the congestion control protocol utilized by the flows (the end-users) and the queues formed inside the network. It is natural to expect that this product-form steady-state distribution is the limit of the steady-state distributions in the original model under the diffusion scaling. Results of this type are known for certain queueing systems such as generalized Jackson networks; see the work by Gamarnik and Zeevi [22]. On the other hand, the validity of such a steady-state diffusion approximation was not known for the bandwidth-sharing networks, under the unweighted proportionally fair policy ($\alpha = 1$); it will be established in this thesis.

Here we remark that with the sole exception of unweighted proportional fairness

($\alpha = 1$) in bandwidth-sharing networks, the state-of-the-art heavy-traffic analysis described above only establishes the dependence of queue sizes/number of flows on the system load ρ (as $1/(1-\rho)$), and stops short of establishing their dependence on the network structure. In the next chapter, we will see a natural connection between switched networks and bandwidth-sharing networks, which implies that the dependence on the network structure can be established for a certain (randomized) version of proportional fairness in switched networks[1], under certain technical assumptions similar to the local traffic condition in [32]. It is then natural to ask whether, without these assumptions, dependence on the network structure can be established for general switched networks operating under (a version of) proportional fairness, or for that matter, under some other online policy. This question is relevant because there are popular examples, such as input-queued switches, that do not satisfy these technical assumptions. In this regard, we put forth a conjecture (Conjecture 4.7.1) on the performance of proportional fairness in input-queued switches, and propose and analyze an online policy for general switched networks in Chapter 5, where we establish its dependence on the network structure.

Performance vs Complexity Tradeoff. So far we have reviewed literature on the performance aspect of resource allocation policies. As mentioned earlier, besides guaranteeing provably good performance, another important aspect of a policy is its complexity. While not the focus of the thesis, we remark briefly that sometimes low complexity and good performance are not achievable simultaneously; for example, in certain wireless networks (see [49]). For the types of networks considered in this thesis (for example, input-queued switches), it is not entirely clear whether this tradeoff is fundamental, i.e., whether performance (complexity) has to be sacrificed to guarantee low complexity (good performance). For further discussion, see Chapter 7.

[1] This dependence can be captured in the diffusion approximation under proportional fairness in switched networks, similar to the one established in Kang et al. [32] for bandwidth-sharing networks. This is because the fluid models are identical, and the entire machinery of Kang et al. [32], building upon the work of Bramson [9] and Williams [64], relies on a fluid model. For more concrete discussion, see Section 4.7.

1.3 Contributions of The Thesis

Performance Properties of α-Weighted Policies. We refer to the MW-α and α-fair policies collectively as α-weighted policies. We advance the performance analysis of α-weighted policies, in both the steady-state and the transient regimes.

In switched networks, for the transient regime, we obtain a probabilistic bound on the maximal (over a finite time horizon) queue size, when operating under a MW-α policy with $\alpha \geq 1$. This result is obtained by combining a Lyapunov drift inequality with a natural extension of Doob's maximal inequality for non-negative supermartingales. Our probabilistic bound, together with prior results on multiplicative state space collapse, leads immediately to a stronger property, namely, full state space collapse, for the case where $\alpha \geq 1$.

For the steady-state regime, we obtain non-asymptotic and explicit bounds on the tail of the distribution of queue sizes, for any $\alpha > 0$. In the process, we establish that, for any $\alpha > 0$, all moments of the steady-state queue sizes are finite. These results are proved by working with a *normed* version of the Lyapunov function that was used in prior work. Specifically, we establish that this normed version is also a Lyapunov function for the system (i.e., it satisfies a drift inequality). It also happens to be a Lipschitz continuous function and this helps crucially in establishing exponential tail bounds, using results of Hajek [27] and Bertsimas, Gamarnik, and Tsitsiklis [4]. We note that our bounds on the tail exponent depend explicitly on the system load and the total number of queues, in contrast to earlier works [57, 59, 62]. In particular, the bounds take the form $-N^{-\beta}(1-\rho)$, with β a positive constant that depends on α, and where N is the number of queues, and ρ the system load.

In bandwidth-sharing networks, for the transient regime, a probabilistic bound on the maximal (over a finite time horizon) number of flows, when operating under an α-fair policy with $\alpha \geq 1$ is also obtained, using similar techniques to those for switched networks. An immediate corollary of this probabilistic bound is full state space collapse, for the case where $\alpha \geq 1$.

For the steady-state regime, we obtain non-asymptotic and explicit bounds on

the tail of the distribution of the number of flows, for any $\alpha > 0$. In the process, we establish that, for any $\alpha > 0$, all moments of the steady-state number of flows are finite. Techniques used to established these results are similar to those used for switched networks.

For $\alpha = 1$, the exponent in the exponential tail bound that we establish for the distribution of the number of flows is proportional to a suitably defined distance ("gap") from critical loading; this gap is of the same type as the familiar $1 - \rho$ term, where ρ is the usual load factor in a queueing system. This particular dependence on the load leads to the tightness of the steady-state distributions of the model under diffusion scaling. It leads to one important consequence, namely, the validity of the diffusion approximation, in steady state.

Scaling Analysis in Switched Networks. As mentioned earlier, the high-level performance question of interest is identifying optimal scaling behavior of policies in large-scale switched networks, and designing policies that achieve it. More specifically, the performance objective of interest is the long-run average total queue size in the system, and we address the following two questions: (a) what is the minimal value of the performance objective among the class of online policies, and (b) how does it depend on the network structure and the system load.

For a general single-hop switched network, we propose a new online policy, which we call **EMUL**. This policy effectively emulates an insensitive bandwidth-sharing network with a product-form stationary distribution, with each component of this product form behaving like an $M/M/1$ queue. This crisp description of the stationary distribution allows us to obtain precise bounds on the average queue sizes under this policy. This leads to establishing, as a corollary of our result, the validity of a conjecture stated in [51] for input-queued switches. In general, it provides explicit upper bounds on the long-run average total queue size for any switched network. Furthermore, due to the explicit bound on the stationary distribution of queue sizes under our policy, we are able to establish a form of large-deviations optimality of the policy for a class of single-hop switched networks, including input-queued switches.

We would like to note that for input-queued switches, the policy above gives an upper bound of the form $\sqrt{N}/(1-\rho)+N^{1.5}$ on the long-run average total queue size. In the heavy-traffic regime, where $\rho \to 1$, this is a significant improvement over the best known bounds of $O(N/(1-\rho))$ (due to the moment bounds of [42] for the maximum weight policy) or $O(\sqrt{N}\log N/(1-\rho)^2)$ (obtained by using a batching policy [44]). However, in the regime where $\rho = 1 - 1/\sqrt{N}$, $N \to \infty$, all existing bounds, including the one produced by the policy above, give an upper bound $O(N^{1.5})$ (ignoring poly-logarithmic dependence on N) on the long-run average total queue size. In contrast, the conjectured optimal scaling $O\left(\sqrt{N}/(1-\rho)\right)$ gives an upper bound $O(N)$, when $\rho = 1 - 1/\sqrt{N}$. It is then natural to ask whether this gap can be reduced, i.e., whether there exists a policy under which the long-run average total queue size is upper bounded by $O(N^{\beta})$, with $\beta < 1.5$ (and ideally, $\beta = 1$), when $\rho = 1 - 1/\sqrt{N}$, and $N \to \infty$.

In input-queued switches, we propose a new policy under which the long-run average total queue size is upper bounded by $O\big(N^{0.75}f(N)\log f(N)\big)$, in the regime where $\rho = 1 - 1/f(N)$, $f(N) \geq \sqrt{N}$, and $N \to \infty$, and where arrival rates to all queues are uniform. As a corollary, in the same regime, and when $f(N) = \sqrt{N}$, the long-run average total queue size is upper bounded by $O(N^{1.25}\log N)$. This is the best known scaling with respect to N, when $\rho = 1 - 1/\sqrt{N}$. While this is a significant improvement over existing bounds, we believe that the right scaling is $O(N)$[2]. The current best known scalings on the long-run average total queue size in various different regimes, in a general input-queued switch with N queues, are summarized in Table 1.1 (note that an input-queued switch with N queues has $\sqrt{N}$ input ports and $\sqrt{N}$ output ports).

1.4 Organization

The rest of the thesis is organized as follows. In Chapter 2, we describe the models that will be considered throughout the thesis. In particular, we introduce the

[2]For detailed discussion, see Section 6.5.

Table 1.1: Best known scalings for an input-queued switch with N queues and under load factor ρ, in various regimes

Regime	Upper Bound Scaling	Lower Bound Scaling	References
$\frac{1}{1-\rho} < \sqrt{N}$	$O\left(\frac{\sqrt{N}\log N}{(1-\rho)^2}\right)$	$\Omega\left(\frac{\sqrt{N}}{1-\rho}\right)$	[44]
$\frac{1}{1-\rho} = \sqrt{N}$	$O\left(N^{1.25}\log N\right)$	$\Omega\left(N\right)$	this thesis, [50]
$\sqrt{N} \leq \frac{1}{1-\rho} < N$	$O\left(\frac{N^{0.75}\log N}{1-\rho}\right)$	$\Omega\left(\frac{\sqrt{N}}{1-\rho}\right)$	this thesis, [50]
$\frac{1}{1-\rho} \geq N$	$O\left(\frac{\sqrt{N}}{1-\rho}\right)$	$\Omega\left(\frac{\sqrt{N}}{1-\rho}\right)$	this thesis, [52]

switched network model and the bandwidth-sharing network model, and also point out a connection between these models. This connection is reflected indirectly in the similarities between Chapter 3 and Chapter 4, and is used directly in Chapter 5.

The first part of the thesis consists of Chapters 3 and 4, which consider performance properties of α-weighted policies. In Chapter 3, we describe our results on the performance properties of MW-α policies in switched networks, and, in Chapter 4, those of α-fair policies in bandwidth-sharing networks. The methods developed in the two chapters are very similar, hence the reader may wish to skip the technical sections in either chapter.

The second part of the thesis, on scaling analysis in switched networks, consists of Chapters 5 and 6. The policy **EMUL**, which produces optimal heavy-traffic queue-size scaling in a class of switched networks, including input-queued switches, will be described in Chapter 5. Chapter 6 contains the result on queue-size scaling in input-queued switches. We conclude the thesis in Chapter 7 with some discussion and a list of open problems.

Results in Chapter 3 have appeared in [53], results in Chapter 4 have appeared in [54], results in Chapter 5 in [52], and those in Chapter 6 in [50].

Chapter 2

Notation and Models

This chapter details the notation used and the models studied in the thesis. In particular, we describe single-hop switched networks in Section 2.2, input-queued switches, an important example of switched networks, in Section 2.2.1, and bandwidth-sharing networks in Section 2.3. We point out an important connection between switched networks and bandwidth-sharing networks in Section 2.4.

2.1 Notation

Let $\mathbb{Z}$ be the set of integers, $\mathbb{N}$ the set of natural numbers $\{1, 2, \dots\}$, and $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$. Let $\mathbb{R}$ be the set of real numbers, $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geq 0\}$, and $\mathbb{R}_p = \{x \in \mathbb{R} : x > 0\}$. Let $\mathbb{Z}^N$ the integer lattice space of dimension N, and let $\mathbb{Z}_+^N \subset \mathbb{Z}^N$ consist of points with only nonnegative components. Let $\mathbb{R}^N$ denote the real vector space of dimension N, $\mathbb{R}_+^N$ the nonnegative orthant of $\mathbb{R}^N$, and $\mathbb{R}_p^N$ the set of all real vectors in $\mathbb{R}^N$ with positive components. For $x \in \mathbb{R}$, let $[x]^+ = \max(x, 0)$. When $\mathbf{x}$ is a vector, the maximum is taken componentwise. We reserve bold letters for real vectors and matrices. For example, we write $\mathbf{x} = [x_i]_{1 \leq i \leq N}$ for a typical vector in $\mathbb{R}^N$. The vector $\mathbf{e}_i$ is the ith unit vector, with all components being 0 except for the ith component, which is equal to 1. The vector $\mathbf{0}$ is the vector with all components being 0. $\mathbf{1}$ denotes the vector of all 1s. The indicator function of an event A is denoted by $\mathbb{I}_A$. For a natural number m, $m! = \prod_{i=1}^{m} i$ is the factorial of m, and

by convention, $0! = 1$. For a probability distribution $\boldsymbol{\pi}$, $\mathbb{E}_{\boldsymbol{\pi}}[\cdot]$ and $\mathbb{P}_{\boldsymbol{\pi}}[\cdot]$ denote the expectation and probability taken with respect to the distribution $\boldsymbol{\pi}$, respectively. We use the shorthand "iff" for the phrase "if and only if".

2.2 Switched Networks (SN)

The Model. We consider the following *single-hop* switched network model, following closely the exposition in [55].

Consider a collection of N queues. Let time be discrete: time slot $\tau \in \{0, 1, \ldots\}$ runs from time τ to $\tau + 1$. Let $Q_i(\tau)$ denote the (nonnegative integer) length of queue $i \in \{1, 2, \ldots, N\}$ at the beginning of time slot τ, and let $\mathbf{Q}(\tau)$ be the vector $(Q_i(\tau))_{i=1}^N$, so that the vector of initial queue lengths is $\mathbf{Q}(0)$. Unit-sized *packets* arrive to the system over time, and let $A_i(\tau)$ denote the total number of packets that arrive to queue i up to the beginning of time slot τ, so that $A_i(0) = 0$ for all $i \in \{1, 2, \ldots, N\}$. Let $\mathbf{A}(\tau) = (A_i(\tau))_{i=1}^N$ be the vector of cumulative arrivals, and let $a_i(\tau) = A_i(\tau+1) - A_i(\tau)$ be the number of packets that arrive to queue i during time slot τ.

At the very beginning of time slot τ, the queue vector $\mathbf{Q}(\tau)$ is offered service described by a vector $\boldsymbol{\sigma}(\tau) = (\sigma_i(\tau))_{i=1}^N$ drawn from a given finite set $\mathcal{S} \subset \mathbb{R}_+^N$ of *feasible schedules*. A resource allocation policy in the context of switched networks will be called a *scheduling policy*, and decides which schedule to use in each time slot. Throughout this thesis, we will only consider *online* policies; that is, the scheduling decision in time slot τ will be based only on historical information, i.e., the arrival process $\mathbf{A}(\cdot)$ till the beginning of time slot τ.

In this thesis, we focus on the special case $\mathcal{S} \subset \{0, 1\}^N$. Note that this is not a very restrictive assumption; we make this assumption for ease of exposition and for illustration of ideas. Let $S_i(\tau) = \sum_{\tau'=0}^{\tau-1} \sigma_i(\tau')$ be the total service offered to queue i, up to the end of time slot $\tau - 1$.

Given the above description, the queues evolve according to the relation

$$Q_i(\tau+1) = \left[Q_i(\tau) - \sigma_i(\tau)\right]^+ + a_i(\tau), \tag{2.1}$$

for each $i \in \{1, 2, \ldots, N\}$. We define $dZ_i(\tau) = \left[\sigma_i(\tau) - Q_i(\tau)\right]^+$, which is the amount of idling at queue i in time slot τ, define $Z_i(0) = 0$, and let $Z_i(\tau) = \sum_{\tau'=0}^{\tau-1} dZ_i(\tau')$, for $\tau > 0$. Then $Z_i(\tau)$ is the cumulative amount of idling at queue i, up to the beginning of time slot τ. Hence, (2.1) can be also written as

$$Q_i(\tau) = Q_i(0) + A_i(\tau) - S_i(\tau) + Z_i(\tau). \tag{2.2}$$

In order to avoid trivialities, we assume, throughout the thesis, the following.

Assumption 2.2.1 *For every queue i, there exists a $\boldsymbol{\sigma} \in \mathcal{S}$ such that $\sigma_i = 1$.*

A Note on Arrival Processes. In this thesis, we will make different assumptions on the arrival processes in a **SN**. More specifically, in Chapters 3 and 6, we assume that the arrival processes $A_i(\cdot)$ are independent Bernoulli processes with arrival rates λ_i, so that $a_i(\tau) \in \{0, 1\}$, $\mathbb{E}[a_i(\tau)] = \lambda_i$ for all i and τ. In Chapter 5, we assume that the arrival processes $A_i(\cdot)$ are independent Poisson processes with arrival rates λ_i, so that in particular, $a_i(\tau)$ is a Poisson random variable with mean $1/\lambda_i$ for each i and τ. We remark that these assumptions are made to simplify the analysis and highlight the core ideas, and hence not very restrictive. For example, results similar to those stated in Chapter 3 hold under the assumption that the arrivals are Poisson. In general, key features of the arrival processes which we require for the analysis to go through, are that $a_i(\tau)$ are i.i.d random variables for all i and τ, and that $a_i(\tau)$ are sufficiently light-tailed.

If the arrival process $A_i(\cdot)$ has an arrival rate λ_i, $i \in \{1, 2, \ldots, N\}$, then we call $\boldsymbol{\lambda} = (\lambda_i)_{i=1}^N$ the *arrival rate vector* for the system.

Admissible Region and Load.

Definition 2.2.2 (Admissible region) *Let $\mathcal{S} \subset \{0,1\}^N$ be the set of allowed schedules. Let $\langle \mathcal{S} \rangle$ be the convex hull of $\mathcal{S}$, i.e.,*

$$\langle \mathcal{S} \rangle = \Big\{ \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} \boldsymbol{\sigma} : \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} = 1, \text{ and } \alpha_{\boldsymbol{\sigma}} \geq 0 \text{ for all } \boldsymbol{\sigma} \Big\}. \tag{2.3}$$

We define the admissible region $\bar{\boldsymbol{\Lambda}}$ *to be*

$$\bar{\boldsymbol{\Lambda}} = \big\{ \boldsymbol{\lambda} \in \mathbb{R}_+^N : \boldsymbol{\lambda} \leq \boldsymbol{\sigma} \text{ componentwise, for some } \boldsymbol{\sigma} \in \langle \mathcal{S} \rangle \big\}. \tag{2.4}$$

We also define the strictly admissible region $\boldsymbol{\Lambda}$ *to be the interior of the admissible region $\bar{\boldsymbol{\Lambda}}$, which can also be written as*

$$\boldsymbol{\Lambda} = \big\{ \boldsymbol{\lambda} \in \mathbb{R}_+^N : \boldsymbol{\lambda} < \boldsymbol{\sigma} \text{ componentwise, for some } \boldsymbol{\sigma} \in \langle \mathcal{S} \rangle \big\}. \tag{2.5}$$

Definition 2.2.3 (Static planning problems and load) *Define the static planning optimization problem* PRIMAL($\boldsymbol{\lambda}$) *for $\boldsymbol{\lambda} \in \mathbb{R}_+^N$ to be*

$$\begin{aligned} \textit{minimize} \quad & \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} && (2.6) \\ \textit{subject to} \quad & \boldsymbol{\lambda} \leq \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} \boldsymbol{\sigma}, && (2.7) \\ & \alpha_{\boldsymbol{\sigma}} \in \mathbb{R}_+, \text{ for all } \boldsymbol{\sigma} \in \mathcal{S}. && (2.8) \end{aligned}$$

Define the load induced by $\boldsymbol{\lambda}$, denoted by $\rho_{SN}(\boldsymbol{\lambda})$ (and, when the context is clear, $\rho(\boldsymbol{\lambda})$), as the value of the optimization problem PRIMAL($\boldsymbol{\lambda}$).

Note that $\boldsymbol{\lambda}$ is (strictly) admissible if and only if $\rho(\boldsymbol{\lambda}) \leq 1$ ($\rho(\boldsymbol{\lambda}) < 1$). Note also that $\boldsymbol{\lambda}$ is strictly admissible, i.e. $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$, iff there is a policy under which the Markov chain describing the network is positive recurrent. For that reason, the strictly admissible region $\boldsymbol{\Lambda}$ is often called the capacity region as well.

The following is a simple and useful property of $\rho(\cdot)$: for any $\mathbf{a}$, $\mathbf{b} \in \mathbb{R}_+^N$,

$$\rho(\mathbf{a} + \mathbf{b}) \leq \rho(\mathbf{a}) + \rho(\mathbf{b}). \tag{2.9}$$

2.2.1 Input-Queued Switches

An important special case of the switched network model is the so-called input-queued switch, which describes the *switching* mechanism in modern Internet routers. An Internet router has several input ports and output ports. Data packets arrive to the input ports, and are then transferred to the correct output ports through a mechanism known as *switching.*

In general, an $n \times n$ input-queued switch has n input ports and n output ports. It has a separate queue for each input-output pair (i, j), denoted by Q_{ij}[1], and hence in total $N = n^2$ queues. The switch operates in discrete time. At each time slot, the switch fabric can transmit a number of packets from input ports to output ports, subject to the *matching* constraints: each input can transmit at most one packet, and each output can receive at most one packet.

The corresponding schedule set $\mathcal{S}$ is defined as

$$\mathcal{S} = \Big\{ \boldsymbol{\sigma} \in \{0,1\}^{n \times n} : \sum_{m=1}^{n} \sigma_{k,m} \leq 1, \ \sum_{m=1}^{n} \sigma_{m,\ell} \leq 1, \ \ 1 \leq k, \ell \leq n \Big\}. \tag{2.10}$$

The admissible region $\bar{\mathbf{\Lambda}}$ is given by

$$\bar{\mathbf{\Lambda}} = \Big\{ \mathbf{x} \in [0,1]^{n \times n} : \sum_{m=1}^{n} x_{k,m} \leq 1, \ \sum_{m=1}^{n} x_{m,\ell} \leq 1, \ \ 1 \leq k, \ell \leq n \Big\}. \tag{2.11}$$

Finally, for an arrival rate matrix[2] $\boldsymbol{\lambda} \in [0,1]^{n \times n}$, $\rho(\boldsymbol{\lambda})$ is given by

$$\rho(\boldsymbol{\lambda}) = \max_{1 \leq k, \ell \leq n} \Big\{ \sum_{m=1}^{n} \lambda_{k,m}, \sum_{m=1}^{n} \lambda_{m,\ell} \Big\}. \tag{2.12}$$

[1]Here we deviate from our convention of indexing queues by a single subscript. This will ease exposition in the context of input-queued switches, without causing confusion.

[2]Not a vector, for notational convenience, as discussed in the previous footnote.

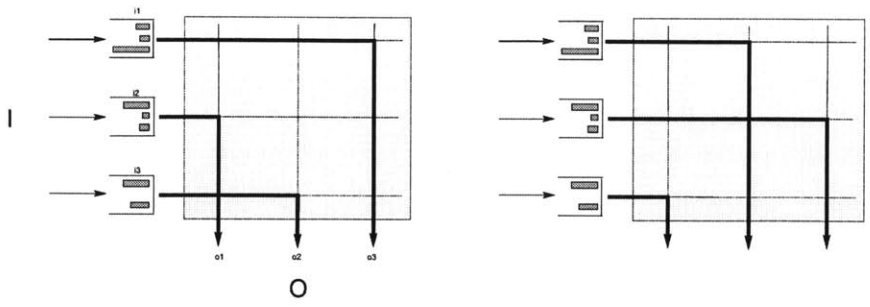


Figure 2-1: An input-queued switch, and two example matchings of inputs to outputs.

2.3 Bandwidth-Sharing Networks (BN)

The Model. Here we describe the bandwidth-sharing network model. We will follow [36] closely, adapting notation accordingly, for the purpose of this thesis. As explained in detail in [36], this model faithfully captures the long-term (or macro level) behavior of congestion control in the current Internet.

Let time be continuous and indexed by $t \in \mathbb{R}_+$. Consider a network with a finite set $\mathcal{J}$ of resources and a set $\mathcal{I}$ of routes, where a route is identified with a non-empty subset of the resource set $\mathcal{J}$. Let $J = |\mathcal{J}|$, and $N = |\mathcal{I}|$. Let $\mathbf{R}$ be a $J \times N$ matrix with nonnegative entries $R_{ji} \geq 0$, where R_{ji} is to be interpreted as the amount of resource j consumed by a unit amount of work on route i. We call $\mathbf{R}$ the *incidence matrix* associated with the **BN**. Let $\mathbf{C} = (C_j)_{j \in \mathcal{J}}$ be a *capacity* vector, where we assume that each entry C_j is a given positive constant. Let the number of flows on route i at time t be denoted by $M_i(t)$, and define the flow vector at time t by $\mathbf{M}(t) = (M_i(t))_{i \in \mathcal{I}}$. Each arriving flow brings in a certain amount of work, which receives service from the network according to a *bandwidth allocation policy.* Let $\phi_i(t)$ be the total *bandwidth* allocated to flows on route i at time t, so that if there are M_i flows on route i, then each flow on route i gets a service rate $\phi_i(t)/M_i$ if $M_i > 0$ (and 0 if $M_i = 0$). Once a flow is served, it departs the network. The capacity constraints are

$$\mathbf{R}\boldsymbol{\phi}(t) \leq \mathbf{C}, \quad \text{for all time } t,$$

where $\boldsymbol{\phi}(t) = (\phi_i(t))_{i\in\mathcal{I}}$. A bandwidth allocation $\boldsymbol{\phi}(t)$ is called *admissible* if the capacity constraints are satisfied.

Arrival Processes and Service Requirement Distributions. Throughout this thesis, we assume that for each route i, new flows arrive as an independent Poisson process of rate ν_i. In Chapter 4, we assume that each arriving flow brings an amount of work (data that it wishes to transfer) which is an exponentially distributed random variable with mean $1/\mu_i$, independent of everything else. In Chapter 5, we will consider a different assumption on the work brought in by arriving flows. There, we will assume that each flow brings a unit amount of work deterministically. Under this latter assumption, there exists a natural connection between **SN** and **BN**, which will be explained in the following section.

Bandwidth Allocation Policy. As mentioned above, a resource allocation policy in the context of bandwidth-sharing networks is called a *bandwidth allocation policy.* In this thesis, we will only consider *online, myopic* policies. A policy is *myopic* if for each i, the total bandwidth $\phi_i(t)$ allocated to route i will only depend on the flow vector $\mathbf{M}(t)$, and we can write $\boldsymbol{\phi}(t) = \boldsymbol{\phi}(\mathbf{M}(t))$.

Admissible Region and Load. For each $i \in \mathcal{I}$, flows of type i bring to the system an average of $\lambda_i = \nu_i/\mu_i$ units of work per unit time. We define the *admissible region* to be

$$\bar{\boldsymbol{\Lambda}} = \{\mathbf{x} \in \mathbb{R}_+^N : \mathbf{R}\mathbf{x} \leq \mathbf{C}\}.$$

For any policy, in order for the Markov process describing the network to be positive recurrent, it is necessary that

$$\mathbf{R}\boldsymbol{\lambda} < \mathbf{C}, \text{ componentwise.} \tag{2.13}$$

We note that under the α-fair bandwidth-sharing policy, Condition (2.13) is also sufficient for positive recurrence of the process $\mathbf{M}(\cdot)$ [6, 15, 36]. Similar to switched

networks, we define the *strictly admissible region* to be $\mathbf{\Lambda} = \{\mathbf{x} \in \mathbb{R}_+^N : \mathbf{Rx} < \mathbf{C}\}$, and a vector $\boldsymbol{\lambda} \in \mathbf{\Lambda}$ is called *strictly admissible.*

For a vector $\boldsymbol{\lambda} \in \mathbb{R}_+^N$, define the load $\rho_{\mathbf{BN}}(\boldsymbol{\lambda})$ induced by $\boldsymbol{\lambda}$ as follows (and, similar to the case of **SN**, we denote the load as $\rho(\boldsymbol{\lambda})$ when the context is clear). For each $j \in \mathcal{J}$, let $\widetilde{\rho}_j = (\mathbf{R}\boldsymbol{\lambda})_j / C_j$, and define

$$\rho(\boldsymbol{\lambda}) = \max_{j \in \mathcal{J}} \widetilde{\rho}_j.$$

As in Section 2.2, $\boldsymbol{\lambda}$ is (strictly) admissible iff $\rho(\boldsymbol{\lambda}) \leq 1$ ($\rho(\boldsymbol{\lambda}) < 1$). In a sense to be made precise in the following section (Lemma 2.4.1), the definitions of load in the context of **SN** and **BN** coincide.

2.4 The Relation between SN and BN

The Equivalence of $\rho_{\mathbf{SN}}$ and $\rho_{\mathbf{BN}}$. The models **SN** and **BN** described in this chapter are closely related. First, the concepts of load in the two models are equivalent, in the following sense. Consider a **SN** with schedule set $\mathcal{S}$, and admissible region $\bar{\mathbf{\Lambda}}$. $\bar{\mathbf{\Lambda}}$ can be represented by a polytope of the form

$$\bar{\mathbf{\Lambda}} = \left\{\mathbf{x} \in [0,1]^N : \mathbf{Rx} \leq \mathbf{C}\right\},$$

where $\mathbf{R}$ is a $J \times N$ matrix, all entries R_{ji} of $\mathbf{R}$ are non-negative, and $\mathbf{C} = (C_j)_{j=1}^J$ has $C_j > 0$, for each j. Now consider a **BN** with N routes, corresponding to queues in **SN**, and the same admissible region $\bar{\mathbf{\Lambda}}$ as in the **SN** described above, i.e.,

$$\bar{\mathbf{\Lambda}} = \left\{\mathbf{x} \in [0,1]^N : \mathbf{Rx} \leq \mathbf{C}\right\}.$$

For the **SN** and **BN** considered here, suppose that we are given the same arrival rate vector $\boldsymbol{\lambda} \in \mathbb{R}_+^N$. We can define the load $\rho_{\mathbf{SN}}(\boldsymbol{\lambda})$ induced by $\boldsymbol{\lambda}$ in **SN**, and the load $\rho_{\mathbf{BN}}(\boldsymbol{\lambda})$ in **BN**. The next lemma establishes that $\rho_{\mathbf{SN}}(\boldsymbol{\lambda}) = \rho_{\mathbf{BN}}(\boldsymbol{\lambda})$.

Lemma 2.4.1 *Consider the **SN** and **BN** described above, where the **SN** has schedule*

set $\mathcal{S}$, and both networks have admissible region $\bar{\Lambda}$ given by

$$\bar{\Lambda} = \left\{ \mathbf{x} \in [0,1]^N : \mathbf{R}\mathbf{x} \leq \mathbf{C} \right\}.$$

Then, for any $\boldsymbol{\lambda} \in \mathbb{R}_+^N$, $\rho_{SN}(\boldsymbol{\lambda}) = \rho_{BN}(\boldsymbol{\lambda})$.

Proof. The proof is fairly straightforward, and we include it here for completeness. Let $\boldsymbol{\lambda} \in \mathbb{R}_+^N$. By scaling the rows of $\mathbf{R}$ appropriately, we can assume that $C_j = 1$ for all $j \in \mathcal{J}$. For each $j \in \mathcal{J}$, let $\widetilde{\rho}_j = (\mathbf{R}\boldsymbol{\lambda})_j$. Then $\rho_{\mathbf{BN}}(\boldsymbol{\lambda}) = \max_{j \in \mathcal{J}} \widetilde{\rho}_j$. Also note that for each $\boldsymbol{\sigma} \in \mathcal{S}$, $\mathbf{R}\boldsymbol{\sigma} \leq \mathbf{1}$ componentwise, where $\mathbf{1}$ is the vector of all 1s. First we show that $\rho_{\mathbf{BN}}(\boldsymbol{\lambda}) \leq \rho_{\mathbf{SN}}(\boldsymbol{\lambda})$. Let $(\alpha_{\boldsymbol{\sigma}})_{\boldsymbol{\sigma} \in \mathcal{S}}$ be an optimal solution of the linear program PRIMAL($\boldsymbol{\lambda}$) defined in Section 2.2. Recall that PRIMAL($\boldsymbol{\lambda}$) was defined to be

$$\begin{aligned} \text{minimize} \quad & \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} && (2.14) \\ \text{subject to} \quad & \boldsymbol{\lambda} \leq \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} \boldsymbol{\sigma}, && (2.15) \\ & \alpha_{\boldsymbol{\sigma}} \in \mathbb{R}_+, \text{ for all } \boldsymbol{\sigma} \in \mathcal{S}. && (2.16) \end{aligned}$$

Since $\boldsymbol{\lambda} \leq \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} \boldsymbol{\sigma}$, we have that

$$\mathbf{R}\boldsymbol{\lambda} \leq \mathbf{R}\left(\sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} \boldsymbol{\sigma} \right) = \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} (\mathbf{R}\boldsymbol{\sigma}) \leq \left(\sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} \right) \mathbf{1} = \rho_{\mathbf{SN}}(\boldsymbol{\lambda}) \mathbf{1}.$$

Hence, for each $j \in \mathcal{J}$, $\widetilde{\rho}_j = (\mathbf{R}\boldsymbol{\lambda})_j \leq \rho_{\mathbf{SN}}(\boldsymbol{\lambda})$, and so $\rho_{\mathbf{BN}}(\boldsymbol{\lambda}) \leq \rho_{\mathbf{SN}}(\boldsymbol{\lambda})$.

Next we show that $\rho_{\mathbf{BN}}(\boldsymbol{\lambda}) \geq \rho_{\mathbf{SN}}(\boldsymbol{\lambda})$. If $\rho_{\mathbf{BN}}(\boldsymbol{\lambda}) > 0$, then consider $\widetilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda} / \rho_{\mathbf{BN}}(\boldsymbol{\lambda})$. It is easy to see that $\rho_{\mathbf{BN}}(\widetilde{\boldsymbol{\lambda}}) = 1$. In particular, $\mathbf{R}\widetilde{\boldsymbol{\lambda}} \leq \mathbf{1}$, so $\widetilde{\boldsymbol{\lambda}} \in \bar{\Lambda}$, and by definition, there exist $\alpha_{\boldsymbol{\sigma}} \geq 0$, $\boldsymbol{\sigma} \in \mathcal{S}$, such that $\sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} = 1$, and $\widetilde{\boldsymbol{\lambda}} \leq \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} \boldsymbol{\sigma}$. This means that $\rho_{\mathbf{SN}}(\widetilde{\boldsymbol{\lambda}}) \leq 1$, by the definition of PRIMAL($\widetilde{\boldsymbol{\lambda}}$). Thus,

$$\rho_{\mathbf{SN}}(\widetilde{\boldsymbol{\lambda}}) = \rho_{\mathbf{SN}}\left(\frac{\boldsymbol{\lambda}}{\rho_{\mathbf{BN}}(\boldsymbol{\lambda})} \right) = \frac{\rho_{\mathbf{SN}}(\boldsymbol{\lambda})}{\rho_{\mathbf{BN}}(\boldsymbol{\lambda})} \leq 1,$$

and $\rho_{\mathbf{SN}}(\boldsymbol{\lambda}) \leq \rho_{\mathbf{BN}}(\boldsymbol{\lambda})$. If $\rho_{\mathbf{BN}}(\boldsymbol{\lambda}) = 0$, we show that $\boldsymbol{\lambda} = \mathbf{0}$, so that $\rho_{\mathbf{SN}}(\boldsymbol{\lambda}) = 0$ as well. Indeed, since $\mathcal{S}$ is finite, $\bar{\boldsymbol{\Lambda}}$ is compact. Thus, for each $i \in \mathcal{I}$, there exists $j \in \mathcal{J}$ such that $R_{ji} > 0$, because if not, then there exists $i \in \mathcal{I}$ where $R_{ji} = 0$ for all $j \in \mathcal{J}$. This implies that the set $\bar{\boldsymbol{\Lambda}} = \{\mathbf{x} \geq \mathbf{0} : \mathbf{R}\mathbf{x} \leq \mathbf{1}\}$ contains all points of the form $\beta \mathbf{e}_i$, $\beta \geq 0$, so $\bar{\boldsymbol{\Lambda}}$ is not compact, and hence contradiction. Now for each $i \in \mathcal{I}$, pick $j \in \mathcal{J}$ such that $R_{ji} > 0$, then $R_{ji}\lambda_i \leq \sum_{i \in \mathcal{I}} R_{ji}\lambda_i = 0$, implying that $\lambda_i = 0$. Since i is arbitrary, $\lambda_i = 0$ for all $i \in \mathcal{I}$. □

Given the equivalence between $\rho_{\mathbf{SN}}$ and $\rho_{\mathbf{BN}}$, we will drop the subscripts in the sequel. We will not distinguish these two concepts in Chapter 5, where we explicitly consider both **SN** and a corresponding **BN**. Thus in Chapter 5, for a vector $\boldsymbol{\lambda} \in \mathbb{R}_+^N$, we will use the term *load induced by* $\boldsymbol{\lambda}$ and the notation $\rho(\boldsymbol{\lambda})$ to refer to the same quantity, $\rho_{\mathbf{SN}}(\boldsymbol{\lambda})$ and $\rho_{\mathbf{BN}}(\boldsymbol{\lambda})$.

A Correspondence between SN and BN. We have already seen how **SN** and **BN** can be naturally related. Here we explain this relation in more detail. At a high level, a **BN** can be viewed as a continuous-time analogue of a **SN** (and a **SN**, a discrete-time analogue of a **BN**). More precisely, consider a **SN** with N queues and schedule set $\mathcal{S} \subset \{0,1\}^N$. We assume that the schedule set $\mathcal{S}$ is *monotone*, i.e. if $\boldsymbol{\sigma} \in \mathcal{S}$, and $\boldsymbol{\sigma}' \in \{0,1\}^N$ satisfies $\boldsymbol{\sigma}' \leq \boldsymbol{\sigma}$ componentwise, then $\boldsymbol{\sigma}' \in \mathcal{S}$ as well. Note that this is not a restrictive assumption; for example, it is satisfied in both the input-queued switch model, and the independent-set model of a wireless network.

Under the monotonicity assumption, we have the following simplification. The convex hull $\langle \mathcal{S} \rangle$ of $\mathcal{S}$ and the admissible region $\bar{\boldsymbol{\Lambda}}$ coincide, and can be represented by a polytope of the form

$$\langle \mathcal{S} \rangle = \bar{\boldsymbol{\Lambda}} = \Big\{ \mathbf{x} \in [0,1]^N : \mathbf{R}\mathbf{x} \leq \mathbf{C} \Big\}.$$

Schedules in $\mathcal{S}$ are precisely the extreme points of this polytope.

Now consider a **BN** with N routes, corresponding to the queues in **SN**, and the

admissible region $\bar{\mathbf{\Lambda}}$ given by

$$\bar{\mathbf{\Lambda}} = \Big\{ \mathbf{x} \in [0,1]^N : \mathbf{R}\mathbf{x} \leq \mathbf{C} \Big\}.$$

We also suppose that flows arriving to **BN** have size deterministically 1, which are analogous to unit-sized packets in **SN**. As we can see, this **BN** corresponds naturally to the **SN**. Two key differences between the two networks are: (a) **BN** operates in continuous time, whereas **SN**, in discrete time; and (b) bandwidth allocations in **BN** can utilize all points in the polytope $\bar{\mathbf{\Lambda}}$, whereas schedules in **SN** can only utilize extreme points of this polytope.

In the sequel, the close relation between **SN** and **BN** will be used in Section 4.7, Chapter 4, to motivate a conjecture regarding an optimal scheduling policy in input-queued switches, and more extensively in Chapter 5, for the design of a scheduling policy for a general **SN**, with attractive performance measures.

Chapter 3

Performance of Maximum-Weight-α Policies in SN

In this chapter, we establish various qualitative performance bounds for so-called Maximum-Weight-α (MW-α) policies in switched networks. We first define the MW-α policies in Section 3.1, and then state formally our main results in Section 3.2. In Section 3.3, we present a transient analysis of the MW-α policies, for $\alpha \geq 1$. We start by proving a general lemma, Lemma 3.3.2, which is then specialized to a maximal inequality under the MW-α policy, for $\alpha \geq 1$ (Theorem 3.2.1). We then apply the maximal inequality to prove the full state space collapse result for $\alpha \geq 1$ (Theorem 3.3.8). In Section 3.4, we present the exponential upper bound on the tail probability of the steady-state distribution under the MW-α policy, for $\alpha \in (0, \infty)$. We start by establishing a drift inequality, Theorem 3.4.3, for a suitably defined "normed" Lyapunov function (Definition 3.4.1). This drift inequality is crucial for proving the exponential upper bound, Theorem 3.2.2. We conclude the chapter in Section 3.5 with a discussion of the tightness of our exponential bound, in the context of input-queued switches.

The prerequisite for reading this chapter is the description of the switched network model in Section 2.2.

3.1 Maximum-Weight-α Policies

We now describe the so-called *Maximum-Weight-α* (MW-α) policies. For $\alpha > 0$, we use $\mathbf{Q}^\alpha(\tau)$ to denote the vector $(Q_i^\alpha(\tau))_{i=1}^N$. We define the *weight* of schedule $\boldsymbol{\sigma} \in \mathcal{S}$ to be $\boldsymbol{\sigma} \cdot \mathbf{Q}^\alpha(\tau)$. The MW-$\alpha$ policy chooses, at each time slot τ, a schedule with the largest weight (breaking ties arbitrarily). Formally, during time slot τ, the policy chooses a schedule $\boldsymbol{\sigma}(\tau)$ that satisfies

$$\boldsymbol{\sigma}(\tau) \cdot \boldsymbol{Q}^\alpha(\tau) = \max_{\boldsymbol{\sigma} \in \mathcal{S}} \boldsymbol{\sigma} \cdot \mathbf{Q}^\alpha(\tau).$$

We define the maximum α-weight of the queue length vector $\mathbf{Q}$ by

$$w_\alpha(\mathbf{Q}) = \max_{\boldsymbol{\sigma} \in \mathcal{S}} \boldsymbol{\sigma} \cdot \boldsymbol{Q}^\alpha.$$

When $\alpha = 1$, the policy is simply called the MW policy, and we use the notation $w(\mathbf{Q})$ instead of $w_1(\mathbf{Q})$. We take note of the fact that under the MW-α policy, the resulting Markov chain is known to be positive recurrent, for any $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ (cf. [41]). Recall from Section 2.2 that in this chapter, we will assume that the arrival process to queue i is an independent Bernoulli process with parameter λ_i, $i = 1, 2, \ldots, N$.

3.2 Summary of Results

In this section, we summarize our main results for both the transient and the steady-state regime. The proofs are given in subsequent sections.

3.2.1 Transient Regime

Here we provide a simple inequality on the maximal excursion of the queue-size over a finite time interval, under the MW-α policy, with $\alpha \geq 1$.

Theorem 3.2.1 *Consider a switched network operating under the MW-α policy with $\alpha \geq 1$, and assume that $\rho = \rho(\boldsymbol{\lambda}) < 1$. Suppose that $\mathbf{Q}(0) = \mathbf{0}$. Let $Q_{\max}(\tau) =$*

$\max_{i\in\{1,\ldots,N\}} Q_i(\tau)$, *and* $Q^*_{\max}(T) = \max_{\tau\in\{0,1,\ldots,T\}} Q_{\max}(\tau)$. *Then, for any* $b > 0$,

$$\mathbb{P}\left(Q^*_{\max}(T) \geq b\right) \leq \frac{K(\alpha, N)T}{(1-\rho)^{\alpha-1}b^{\alpha+1}}, \tag{3.1}$$

for some positive constant $K(\alpha, N)$ *depending only on* α *and* N.

As an important application, we use Theorem 3.2.1 to prove a full state space collapse result, for $\alpha \geq 1$, in Section 3.3.3. The precise statement can be found in Theorem 3.3.8. Theorem 3.3.8 resolves Conjecture 7.2 in [55], in the special case of single-hop networks with Bernoulli arrival processes. However, our analysis easily extends to the more general case where the increments in the arrival process are i.i.d and uniformly bounded, and when the network is multi-hop.

3.2.2 Steady-State Regime

The Markov chain $\mathbf{Q}(\cdot)$ that describes a switched network operating under the MW-α policy is known to be positive recurrent, as long as the system is underloaded, i.e., if $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ or, equivalently, $\rho(\boldsymbol{\lambda}) < 1$. It is not hard to verify that this Markov chain is irreducible and aperiodic. Therefore, there exists a unique stationary distribution, which we will denote by $\boldsymbol{\pi}$. We use $\mathbb{E}_{\boldsymbol{\pi}}$ and $\mathbb{P}_{\boldsymbol{\pi}}$ to denote expectations and probabilities under $\boldsymbol{\pi}$.

Exponential Bound on Tail Probabilities. For the MW-α policy, and for any $\alpha \in (0, \infty)$, we obtain an explicit exponential upper bound on the tail probabilities of the queue sizes, in steady state.

Theorem 3.2.2 *Consider a switched network operating under the MW-*α *policy, and assume that* $\rho = \rho(\boldsymbol{\lambda}) < 1$. *There exist positive constants* B *and* B' *(that depend on* α, N *and* ρ*) such that for all* $\ell \in \mathbb{Z}_+$*:*

(a) if $\alpha \geq 1$, *then*

$$\mathbb{P}_{\boldsymbol{\pi}}\left(\|\mathbf{Q}(\tau)\|_{\alpha+1} > B + 2N^{\frac{1}{\alpha+1}}\ell\right) \leq \left(\frac{1}{1+\frac{1-\rho}{2N}}\right)^{\ell+1};$$

(b) if $\alpha \in (0,1)$, *then*

$$\mathbb{P}_{\pi}\left(\|\mathbf{Q}(\tau)\|_{\alpha+1} > B' + 10N^{\frac{1}{\alpha+1}}\ell\right) \leq \left(\frac{5}{5+\frac{1-\rho}{2N}}\right)^{\ell+1}.$$

3.3 Transient Analysis ($\alpha \geq 1$)

In this section, we prove Theorem 3.2.1. First we present a general maximal lemma (Lemma 3.3.2), which is then specialized to the switched network. In particular, we prove a drift inequality (Lemma 3.3.5) for the Lyapunov function $F(\mathbf{x}) = \frac{1}{\alpha+1}\sum_i x_i^{\alpha+1}$. We combine the drift inequality with the maximal lemma to obtain Theorem 3.2.1, a maximal inequality for the switched network. We then apply the maximal inequality to prove Theorem 3.3.8, full state space collapse for $\alpha \geq 1$.

A Second-Order Mean Value Theorem. We will be making extensive use of the following theorem [3], which we state below for easy reference.

Proposition 3.3.1 *Let* $g : \mathbb{R}^N \to \mathbb{R}$ *be twice continuously differentiable over an open sphere* S *centered at a vector* $\mathbf{x}$. *Then, for any* $\mathbf{y}$ *such that* $\mathbf{x}+\mathbf{y} \in S$, *there exists a* $\theta \in [0,1]$ *such that*

$$g(\mathbf{x}+\mathbf{y}) = g(\mathbf{x}) + \mathbf{y}^T\nabla g(\mathbf{x}) + \frac{1}{2}\mathbf{y}^T H(\mathbf{x}+\theta\mathbf{y})\mathbf{y}, \tag{3.2}$$

where $\nabla g(\mathbf{x})$ *is the gradient of* g *at* $\mathbf{x}$, *and* $H(\mathbf{x})$ *is the Hessian of the function* g *at* $\mathbf{x}$.

3.3.1 The Key Lemma

Our analysis relies on the following lemma:

Lemma 3.3.2 *Let* $(\mathscr{F}_n)_{n\in\mathbb{Z}_+}$ *be a filtration on a probability space. Let* $(X_n)_{n\in\mathbb{Z}_+}$ *be a nonnegative* $\mathscr{F}_n$*-adapted stochastic process that satisfies*

$$\mathbb{E}[X_{n+1} \mid \mathscr{F}_n] \leq X_n + B_n \tag{3.3}$$

where the B_n are nonnegative random variables (not necessarily $\mathscr{F}_n$-adapted) with finite means. Let $X_n^* = \max\{X_0, \ldots, X_n\}$ and suppose that $X_0 = 0$. Then, for any $a > 0$ and any $T \in \mathbb{Z}_+$,

$$\mathbb{P}(X_T^* \geq a) \leq \frac{\sum_{n=0}^{T-1} \mathbb{E}[B_n]}{a}.$$

This lemma is a simple consequence of the following standard maximal inequality for nonnegative supermartingales (see for example, Exercise 4, Section 12.4, of [26]):

Theorem 3.3.3 *Let $(\mathscr{F}_n)_{n\in\mathbb{Z}_+}$ be a filtration on a probability space. Let $(Y_n)_{n\in\mathbb{Z}_+}$ be a nonnegative $\mathscr{F}_n$-adapted supermartingale, i.e., for all n,*

$$\mathbb{E}[Y_{n+1} \mid \mathscr{F}_n] \leq Y_n.$$

Let $Y_T^* = \max\{Y_0, \ldots, Y_T\}$. Then,

$$\mathbb{P}(Y_T^* \geq a) \leq \frac{\mathbb{E}[Y_0]}{a}.$$

Proof. (of Lemma 3.3.2) First note that if we take the conditional expectation on both sides of (3.3), given $\mathscr{F}_n$, we have

$$\begin{aligned} \mathbb{E}[X_{n+1} \mid \mathscr{F}_n] &\leq \mathbb{E}[X_n \mid \mathscr{F}_n] + \mathbb{E}[B_n \mid F_n] \\ &= X_n + \mathbb{E}[B_n \mid F_n]. \end{aligned}$$

Fix $T \in \mathbb{Z}_+$. For any $n \leq T$, define

$$Y_n = X_n + \mathbb{E}\left[\sum_{k=n}^{T-1} B_k \;\middle|\; \mathscr{F}_n\right].$$

Then

$$\begin{aligned} \mathbb{E}[Y_{n+1} \mid \mathscr{F}_n] &= \mathbb{E}[X_{n+1} \mid \mathscr{F}_n] + \mathbb{E}\left[\mathbb{E}\left[\sum_{k=n+1}^{T-1} B_k \;\middle|\; \mathscr{F}_{n+1}\right] \;\middle|\; \mathscr{F}_n\right] \\ &\leq X_n + \mathbb{E}[B_n \mid F_n] + \mathbb{E}\left[\sum_{k=n+1}^{T-1} B_k \;\middle|\; \mathscr{F}_n\right] = Y_n. \end{aligned}$$

Thus, Y_n is an $\mathscr{F}_n$-adapted supermartingale; furthermore, by definition, Y_n is nonnegative for all n. Therefore, by Theorem 3.3.3,

$$\mathbb{P}(Y_T^* \geq a) \leq \frac{\mathbb{E}[Y_0]}{a} = \frac{\mathbb{E}\left[\sum_{k=0}^{T-1} B_k\right]}{a}. \qquad \square$$

But $Y_n \geq X_n$ for all n, since the B_k are nonnegative. Thus

$$\mathbb{P}(X_T^* \geq a) \leq \mathbb{P}(Y_T^* \geq a) \leq \frac{\mathbb{E}\left[\sum_{k=0}^{T-1} B_k\right]}{a}.$$

We have the following corollary of Lemma 3.3.2 in which we take all the B_n equal to the same constant:

Corollary 3.3.4 *Let $\mathscr{F}_n$, X_n and X_n^* be as in Lemma 3.3.2. Suppose that*

$$\mathbb{E}[X_{n+1} \mid \mathscr{F}_n] \leq X_n + B,$$

for all $n \geq 0$, where B is a nonnegative constant. Then, for any $a > 0$ and any $T \in \mathbb{Z}_+$,

$$\mathbb{P}(X_T^* \geq a) \leq \frac{BT}{a}.$$

3.3.2 The Maximal Inequality for Switched Networks

We employ the Lyapunov function

$$F(\mathbf{x}) = \frac{1}{\alpha+1} \sum_{i=1}^{N} x_i^{\alpha+1}, \tag{3.4}$$

to study the MW-α policy. This is the Lyapunov function that was used in [41] to establish positive recurrence of the chain $\mathbf{Q}(\cdot)$ under the MW-α policy. Below we fine-tune the proof in [41] to obtain a more precise bound.

Lemma 3.3.5 *Let $\alpha \geq 1$. For a switched network model operating under the MW-α*

policy with $\rho = \rho(\boldsymbol{\lambda}) < 1$, *we have:*

$$\mathbb{E}\big[F(\mathbf{Q}(\tau+1)) - F(\mathbf{Q}(\tau)) \mid \mathbf{Q}(\tau)\big] \leq \frac{\bar{K}(\alpha, N)}{(1-\rho)^{\alpha-1}}, \tag{3.5}$$

where $\bar{K}(\alpha, N)$ *is a constant depending only on* α *and* N.

Proof. Let $\boldsymbol{\delta}(\tau) = \mathbf{Q}(\tau+1) - \mathbf{Q}(\tau)$. Then by the relation (2.1),

$$\delta_i(\tau) = a_i(\tau) - \sigma_i(\tau)\mathbb{I}_{\{Q_i(\tau)>0\}} \in [-1, 1]. \tag{3.6}$$

By the second-order mean value theorem, there exists $\theta \in [0,1]$ such that

$$\begin{aligned} F(\mathbf{Q}(\tau+1)) - F(\mathbf{Q}(\tau)) &= \frac{1}{\alpha+1}\sum_{i=1}^{N}((Q_i(\tau)+\delta_i(\tau))^{\alpha+1} - Q_i^{\alpha+1}(\tau)) \\ &= \sum_{i=1}^{N} Q_i^{\alpha}(\tau)\delta_i(\tau) + \frac{1}{2}\sum_{i=1}^{N}\alpha(Q_i(\tau)+\theta\delta_i(\tau))^{\alpha-1}\delta_i^2(\tau). \end{aligned}$$

Let us consider the second term on the RHS. We have

$$\begin{aligned} \sum_{i=1}^{N}\alpha(Q_i(\tau)+\theta\delta_i(\tau))^{\alpha-1}\delta_i^2(\tau) &\leq \sum_{i=1}^{N}\alpha(Q_i(\tau)+\theta)^{\alpha-1} \leq \sum_{i=1}^{N}\alpha(Q_i(\tau)+1)^{\alpha-1} \\ &\leq \alpha\sum_{i=1}^{N}(2^{\alpha-1}Q_i^{\alpha-1}(\tau)+1) = \alpha 2^{\alpha-1}\sum_{i=1}^{N}Q_i^{\alpha-1}(\tau) + \alpha N \\ &\leq \alpha 2^{\alpha-1}NQ_{\max}^{\alpha-1}(\tau) + \alpha N. \end{aligned}$$

The third inequality follows because when $Q_i(\tau) \geq 1$,

$$(Q_i(\tau)+1)^{\alpha-1} \leq (2Q_i(\tau))^{\alpha-1} = 2^{\alpha-1}Q_i^{\alpha-1}(\tau),$$

and when $Q_i(\tau) = 0$,

$$(Q_i(\tau)+1)^{\alpha-1} = 1.$$

Let us now consider the term $\sum_{i=1}^{N} Q_i^{\alpha}(\tau)\delta_i(\tau)$, and its conditional expectation

under $\mathbf{Q}(\tau)$. By Eq. (3.6), we have

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{N} Q_i^\alpha(\tau)\delta_i(\tau) \mid \mathbf{Q}(\tau)\right] &= \sum_{i=1}^{N} Q_i^\alpha(\tau)\lambda_i - \sum_{i=1}^{N} Q_i^\alpha(\tau)\sigma_i(\tau) \\ &= \sum_{i=1}^{N} Q_i^\alpha(\tau)\lambda_i - w_\alpha(\mathbf{Q}(\tau)).\end{aligned}$$

Now consider $\sum_{i=1}^{N} Q_i^\alpha(\tau)\lambda_i$. From the definition of $\rho = \rho(\boldsymbol{\lambda})$, there exist constants $\alpha_{\boldsymbol{\sigma}} \geq 0$ such that $\sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha_{\boldsymbol{\sigma}} \leq \rho$, and

$$\boldsymbol{\lambda} \leq \sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha_{\boldsymbol{\sigma}}\boldsymbol{\sigma}.$$

Therefore,

$$\begin{aligned}\sum_i Q_i^\alpha(\tau)\lambda_i = \mathbf{Q}^\alpha(\tau)\cdot\boldsymbol{\lambda} &\leq \sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha_{\boldsymbol{\sigma}}\mathbf{Q}^\alpha(\tau)\cdot\boldsymbol{\sigma} \\ &\leq \sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha_{\boldsymbol{\sigma}} w_\alpha(\mathbf{Q}(\tau)) \leq \rho w_\alpha(\mathbf{Q}(\tau)).\end{aligned}$$

Thus we have

$$\mathbb{E}\left[\sum_{i=1}^{N} Q_i^\alpha(\tau)\delta_i(\tau) \mid \mathbf{Q}(\tau)\right] \leq -(1-\rho)w_\alpha(\mathbf{Q}(\tau)) \leq -(1-\rho)Q_{\max}^\alpha(\tau).$$

Thus, if we combine the inequalities above, we have

$$\begin{aligned}&\mathbb{E}\big[F(\mathbf{Q}(\tau+1)) - F(\mathbf{Q}(\tau))\big|\mathbf{Q}(\tau)\big] \\ &\leq -(1-\rho)Q_{\max}^\alpha(\tau) + \alpha 2^{\alpha-2}NQ_{\max}^{\alpha-1}(\tau) + \frac{\alpha N}{2}. \qquad (3.7)\end{aligned}$$

It is a simple exercise in calculus to see that the RHS of (3.7) is maximized at $Q_{\max}(\tau) = (\alpha-1)2^{\alpha-2}N/(1-\rho)$, giving the maximum value

$$\frac{(\alpha-1)^{\alpha-1}2^{\alpha(\alpha-2)}N^\alpha}{(1-\rho)^{\alpha-1}} + \frac{\alpha N}{2} \leq \frac{(\alpha-1)^{\alpha-1}2^{\alpha(\alpha-2)}N^\alpha + \alpha N/2}{(1-\rho)^{\alpha-1}}.$$

(3.5) is then established by letting

$$\bar{K}(\alpha, N) = (\alpha - 1)^{\alpha-1} 2^{\alpha(\alpha-2)} N^{\alpha} + \frac{\alpha N}{2}. \qquad \square$$

Proof of Theorem 3.2.1. Let $b > 0$. Then

$$\begin{aligned}
\mathbb{P}\left(Q^*_{\max}(T) \geq b\right) &= \mathbb{P}\left(\frac{1}{\alpha+1}\left(Q^*_{\max}(T)\right)^{\alpha+1} \geq \frac{1}{\alpha+1} b^{\alpha+1}\right) \\
&\leq \mathbb{P}\left(\max_{\tau \in \{0,\ldots,T\}} F(\mathbf{Q}(\tau)) \geq \frac{1}{\alpha+1} b^{\alpha+1}\right).
\end{aligned}$$

Now, by Lemma 3.3.5 and Corollary 3.3.4,

$$\begin{aligned}
\mathbb{P}\left(\max_{\tau \in \{0,\ldots,T\}} F(\mathbf{Q}(\tau)) \geq \frac{1}{\alpha+1} b^{\alpha+1}\right) &\leq \frac{(\alpha+1)\bar{K}(\alpha, N)T}{(1-\rho)^{\alpha-1} b^{\alpha+1}} \\
&= \frac{K(\alpha, N)T}{(1-\rho)^{\alpha-1} b^{\alpha+1}},
\end{aligned}$$

where $K(\alpha, N) = (\alpha + 1)\bar{K}(\alpha, N)$.

3.3.3 Full State Space Collapse for $\alpha \geq 1$

Throughout this section, we assume that we are given $\alpha \geq 1$, and correspondingly, the Lyapunov function $F(\boldsymbol{x}) = \frac{1}{\alpha+1} \sum_{i=1}^{N} x_i^{\alpha+1}$. To state the full state space collapse result for $\alpha \geq 1$, we need some preliminary definitions and the statement of the multiplicative state space collapse result.

We will consider a critical arrival rate vector $\boldsymbol{\lambda}$ with $\rho(\boldsymbol{\lambda}) = 1$. More formally, define $\partial\boldsymbol{\Lambda}$ the set of *critical* arrival rate vectors:

$$\partial\boldsymbol{\Lambda} \;=\; \bar{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda} = \left\{\boldsymbol{\lambda} \in \bar{\boldsymbol{\Lambda}} : \rho(\boldsymbol{\lambda}) = 1\right\}.$$

Now consider the linear optimization problem, named DUAL($\boldsymbol{\lambda}$) in [55]:

$$\begin{aligned} \text{maximize} \quad & \boldsymbol{\xi} \cdot \boldsymbol{\lambda} \\ \text{subject to} \quad & \max_{\boldsymbol{\sigma} \in \mathcal{S}} \boldsymbol{\xi} \cdot \boldsymbol{\sigma} \leq 1, \\ & \boldsymbol{\xi} \in \mathbb{R}_+^N. \end{aligned}$$

For $\boldsymbol{\lambda} \in \partial\boldsymbol{\Lambda}$, the optimal value of the objective in DUAL($\boldsymbol{\lambda}$) is 1 (cf. [55]). The set of optimal solutions to DUAL($\boldsymbol{\lambda}$) is a bounded polyhedron, and we let $\mathcal{S}^* = \mathcal{S}^*(\boldsymbol{\lambda})$ be the set of its extreme points.

Fix $\boldsymbol{\lambda} \in \partial\boldsymbol{\Lambda}$. We then consider the optimization problem ALGD(w):

$$\begin{aligned} \text{minimize} \quad & F(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{\xi} \cdot \boldsymbol{x} \geq w_{\boldsymbol{\xi}} \text{ for all } \boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda}), \\ & \boldsymbol{x} \in \mathbb{R}_+^N. \end{aligned}$$

We know from [55] that ALGD(w) has a unique solution. We now define the *lifting map*:

Definition 3.3.6 *Fix some $\boldsymbol{\lambda} \in \partial\boldsymbol{\Lambda}$. The* lifting map $\Delta^{\boldsymbol{\lambda}} : \mathbb{R}_+^{|\mathcal{S}^*(\boldsymbol{\lambda})|} \to \mathbb{R}_+^N$ *maps w to the unique solution to ALGD(w). We also define the* workload map $W^{\boldsymbol{\lambda}} : \mathbb{R}_+^N \to \mathbb{R}_+^{|\mathcal{S}^*(\boldsymbol{\lambda})|}$ *by* $W^{\boldsymbol{\lambda}}(\mathbf{q}) = (\boldsymbol{\xi} \cdot \mathbf{q})_{\boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda})}$.

Fix $\boldsymbol{\lambda} \in \partial\boldsymbol{\Lambda}$. Consider a sequence of switched networks indexed by $r \in \mathbb{N}$, operating under the MW-α policy (recall that $\alpha \geq 1$ here), all with the same number N of queues and feasible schedules. Suppose that $\boldsymbol{\lambda}^r \in \boldsymbol{\Lambda}$ for all r, and that $\boldsymbol{\lambda}^r = \boldsymbol{\lambda} - \boldsymbol{\Gamma}/r$, for some $\boldsymbol{\Gamma} \in \mathbb{R}_+^N$. For simplicity, suppose that all networks start with empty queues. Consider the following central limit scaling,

$$\hat{\mathbf{q}}^r(t) = \mathbf{Q}^r(r^2 t)/r, \tag{3.8}$$

where $\mathbf{Q}^r(\tau)$ is the queue size vector of the rth network at time τ, and where we extend the domain of $\mathbf{Q}^r(\cdot)$ to $\mathbb{R}_+$ by linear interpolation in each interval $(\tau - 1, \tau)$.

We are finally ready to state the multiplicative state space collapse result (Theorem 8.2 in [55]):

Theorem 3.3.7 *Let $\alpha \geq 1$. Fix $T > 0$, and let*

$$\|\mathbf{x}(\cdot)\| = \sup_{i\in\{1,\ldots,N\},0\leq t\leq T} |x_i(t)|.$$

Under the above assumptions, for any $\varepsilon > 0$,

$$\lim_{r\to\infty} \mathbb{P}\left(\frac{\|\hat{\mathbf{q}}^r(\cdot) - \Delta^{\boldsymbol{\lambda}}(W^{\boldsymbol{\lambda}}(\hat{\mathbf{q}}^r(\cdot)))\|}{\|\hat{\mathbf{q}}^r(\cdot)\| \vee 1} < \varepsilon\right) = 1.$$

We now state and prove the full state space collapse result, which settles Conjecture 7.2 in [55], for single-hop networks with Bernoulli arrivals.

Theorem 3.3.8 *Let $\alpha \geq 1$. Under the same assumptions as in Theorem 3.3.7, and for any $\varepsilon > 0$,*

$$\lim_{r\to\infty} \mathbb{P}\left(\|\hat{\mathbf{q}}^r(\cdot) - \Delta^{\boldsymbol{\lambda}}(W^{\boldsymbol{\lambda}}(\hat{\mathbf{q}}^r(\cdot)))\| < \varepsilon\right) = 1.$$

Proof. First note that since $\boldsymbol{\lambda}^r = \boldsymbol{\lambda} - \boldsymbol{\Gamma}/r$, the corresponding loads satisfy $\rho_r \leq 1 - D/r$, for some positive constant $D > 0$. By Theorem 3.2.1, for any $b > 0$,

$$\begin{aligned}\mathbb{P}\left(\max_{\tau\in\{0,1,\ldots,r^2T\}} Q^r_{\max}(\tau) \geq b\right) &\leq \frac{K(\alpha,N)r^2T}{(1-\rho)^{\alpha-1}b^{\alpha+1}} \\ &\leq \frac{K(\alpha,N)r^{1+\alpha}T}{D^{\alpha-1}b^{\alpha+1}}.\end{aligned}$$

Then with $a = b/r$ and under the scaling in (3.8),

$$\mathbb{P}\big(\|\hat{\mathbf{q}}^r(\cdot)\| \geq a\big) \leq \frac{K(\alpha,N)}{D^{\alpha-1}} \cdot \frac{T}{a^{\alpha+1}},$$

for any $a > 0$.

For notational convenience, we write

$$B(r) = \|\hat{\mathbf{q}}^r(\cdot) - \Delta^{\boldsymbol{\lambda}}(W^{\boldsymbol{\lambda}}(\hat{\mathbf{q}}^r(\cdot)))\|.$$

Then, for any $a > 1$,

$$\begin{aligned}\mathbb{P}\big(B(r) \geq \varepsilon\big) &\leq \mathbb{P}\left(\frac{B(r)}{\|\hat{\mathbf{q}}^r(\cdot)\|} > \frac{\varepsilon}{a} \text{ or } \|\hat{\mathbf{q}}^r(\cdot)\| \geq a\right) \\ &\leq \mathbb{P}\left(\frac{B(r)}{\|\hat{\mathbf{q}}^r(\cdot)\|} > \frac{\varepsilon}{a}\right) + \mathbb{P}\big(\|\hat{\mathbf{q}}^r(\cdot)\| \geq a\big).\end{aligned}$$

Note that by Theorem 3.3.7, the first term on the RHS goes to 0 as $r \to \infty$, for any $a > 0$. The second term on the RHS can be made arbitrarily small by taking a sufficiently large. Thus, $\mathbb{P}(B(r) \geq \varepsilon) \to 0$ as $r \to \infty$. This concludes the proof. □

3.4 Steady-State Analysis ($\alpha > 0$)

3.4.1 MW-α policies: A Useful Drift Inequality

The key to many of our results is a *drift inequality* that holds for every $\alpha > 0$ and $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$. In this section, we shall state and prove this inequality. It will be used in Section 3.4.2 to prove Theorem 3.2.2. We remark that similar drift inequalities for the Lyapunov function given by (3.4), which is related to but different from the Lyapunov function defined in this section, have played an important role in establishing positive recurrence (cf. [61]) and multiplicative state space collapse (cf. [55]).

We now define the Lyapunov function that we will employ. For $\alpha \geq 1$, it will be simply the $(\alpha+1)$-norm $\|\mathbf{x}\|_{1+\alpha}$ of a vector $\mathbf{x}$. However, when $\alpha \in (0,1)$, this function has unbounded second derivatives as we approach the boundary of $\mathbb{R}^N_+$. For this reason, our Lyapunov function will be a suitably smoothed version of $\|\cdot\|_{\alpha+1}$.

Definition 3.4.1 *Define $h_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ to be $h_\alpha(r) = r^\alpha$, when $\alpha \geq 1$, and*

$$h_\alpha(r) = \begin{cases} r^\alpha, & \text{if } r \geq 1, \\ (\alpha-1)r^3 + (1-\alpha)r^2 + r, & \text{if } r < 1, \end{cases}$$

when $\alpha \in (0,1)$. Let $H_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ be the antiderivative of h_α, so that $H_\alpha(r) =$

$\int_0^r h_\alpha(s)\,ds$. *The Lyapunov function* $L_\alpha : \mathbb{R}_+^N \to \mathbb{R}_+$ *is defined to be*

$$L_\alpha(\mathbf{x}) = \left[(\alpha+1)\sum_{i=1}^{N} H_\alpha(x_i)\right]^{\frac{1}{\alpha+1}}.$$

We will make heavy use of various properties of the functions h_α, H_α, and L_α, which we summarize in the following lemma. The proof is elementary and is omitted.

Lemma 3.4.2 *Let* $\alpha \in (0,1)$. *The function* h_α *has the following properties:*

(i) it is continuously differentiable with $h_\alpha(0) = 0$, $h_\alpha(1) = 1$, $h'_\alpha(0) = 1$, *and* $h'_\alpha(1) = \alpha$;

(ii) it is increasing and, in particular, $h_\alpha(r) \geq 0$ *for all* $r \geq 0$;

(iii) we have $r^\alpha - 1 \leq h_\alpha(r) \leq r^\alpha + 1$, *for all* $r \in [0,1]$;

(iv) $h'_\alpha(r) \leq 2$, *for all* $r \geq 0$.

Furthermore, from (iii), we also have the following property of H_α*:*

(iii') $r^{\alpha+1} - 2 \leq (\alpha+1)H_\alpha(r) \leq r^{\alpha+1} + 2$ *for all* $r \geq 0$.

We are now ready to state the drift inequality.

Theorem 3.4.3 *Consider a switched network operating under the MW-*α *policy, and assume that* $\rho = \rho(\boldsymbol{\lambda}) < 1$. *Then, there exists a constant* $B > 0$ *(that depends on* α, N *and* ρ*), such that if* $L_\alpha(\mathbf{Q}(\tau)) > B$, *then*

$$\mathbb{E}[L_\alpha(\mathbf{Q}(\tau+1)) - L_\alpha(\mathbf{Q}(\tau)) \mid \mathbf{Q}(\tau)] \leq -\frac{1-\rho}{2} N^{\frac{1}{\alpha+1}-1}. \tag{3.9}$$

The proof of this drift inequality is quite tedious when $\alpha \neq 1$. To make the proof more accessible and to provide intuition, we first present the somewhat simpler proof for $\alpha = 1$. We then provide the proof for the case of general α, by considering separately the two cases where $\alpha > 1$ and $\alpha \in (0,1)$.

We wish to draw attention here to the main difference from related drift inequalities in the literature. The usual proof of stability involves the Lyapunov function

$\|\mathbf{Q}\|_{\alpha+1}^{\alpha+1}$; for instance, for the standard MW policy, it involves a quadratic Lyapunov function. In contrast, we use $\|\mathbf{Q}\|_{\alpha+1}$ (or its smoothed version), which scales linearly along radial directions. In this sense, our approach is similar in spirit to [4], which employed piecewise linear Lyapunov functions to derive drift inequalities and then moment and tail bounds.

Proof of Theorem 3.4.3: $\alpha = 1$. We first consider the case where $\alpha = 1$. As remarked earlier, we have $L_\alpha(\mathbf{x}) = \|\mathbf{x}\|_2$.

Suppose that $\|\mathbf{Q}(\tau)\|_2 > 0$. We claim that on every sample path, we have

$$\|\mathbf{Q}(\tau+1)\|_2 - \|\mathbf{Q}(\tau)\|_2 \leq \frac{\boldsymbol{Q}(\tau)\cdot\boldsymbol{\delta}(\tau) + \|\boldsymbol{\delta}(\tau)\|_2^2}{\|\mathbf{Q}(\tau)\|_2}, \tag{3.10}$$

where $\boldsymbol{\delta}(\tau) = \mathbf{Q}(\tau+1) - \mathbf{Q}(\tau)$. To see this, we proceed as follows. We have

$$\begin{aligned}\left(\|\mathbf{Q}(\tau)\|_2 + \frac{\mathbf{Q}(\tau)\cdot\boldsymbol{\delta}(\tau) + \|\boldsymbol{\delta}(\tau)\|_2^2}{\|\mathbf{Q}(\tau)\|_2}\right)^2 &\geq \|\mathbf{Q}(\tau)\|_2^2 + 2\left(\boldsymbol{Q}(\tau)\cdot\boldsymbol{\delta}(\tau) + \|\boldsymbol{\delta}(\tau)\|_2^2\right)\\ &\geq \|\mathbf{Q}(\tau)\|_2^2 + 2\boldsymbol{Q}(\tau)\cdot\boldsymbol{\delta}(\tau) + \|\boldsymbol{\delta}(\tau)\|_2^2\\ &= \|\mathbf{Q}(\tau)+\boldsymbol{\delta}(\tau)\|_2^2 = \|\mathbf{Q}(\tau+1)\|_2^2. \end{aligned}\tag{3.11}$$

Note that

$$\|\mathbf{Q}(\tau)\|_2^2 + \boldsymbol{Q}(\tau)\cdot\boldsymbol{\delta}(\tau) + \|\boldsymbol{\delta}(\tau)\|_2^2 = \left\|\mathbf{Q}(\tau) + \frac{\boldsymbol{\delta}(\tau)}{2}\right\|_2^2 + \frac{3}{4}\|\boldsymbol{\delta}(\tau)\|_2^2 \geq 0.$$

We divide by $\|\mathbf{Q}(\tau)\|_2$, to obtain

$$\|\mathbf{Q}(\tau)\|_2 + \frac{\boldsymbol{Q}(\tau)\cdot\boldsymbol{\delta}(\tau) + \|\boldsymbol{\delta}(\tau)\|_2^2}{\|\mathbf{Q}(\tau)\|_2} \geq 0.$$

Therefore, we can take square roots of both sides of (3.11), without reversing the direction of the inequality, and the claimed inequality (3.10) follows.

Recall that $|\delta_i(\tau)| \leq 1$, because of the Bernoulli arrival assumption. It follows that $\|\boldsymbol{\delta}(\tau)\|_2 \leq N^{1/2}$. We now take the conditional expectation of both sides of (3.10).

We have

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{Q}(\tau+1)\|_2 - \|\mathbf{Q}(\tau)\|_2 \;\middle|\; \mathbf{Q}(\tau)\right] &\leq \mathbb{E}\left[\frac{\boldsymbol{Q}(\tau)\cdot\mathbf{a}(\tau) - \mathbf{Q}(\tau)\cdot\boldsymbol{\sigma}(\tau) + N}{\|\mathbf{Q}(\tau)\|_2} \;\middle|\; \mathbf{Q}(\tau)\right] \\
&= \frac{\sum_{i=1}^N Q_i(\tau)\mathbb{E}\left[a_i(\tau)\right] - \mathbf{Q}(\tau)\cdot\boldsymbol{\sigma}(\tau) + N}{\|\mathbf{Q}(\tau)\|_2} \\
&= \frac{\sum_{i=1}^N Q_i(\tau)\lambda_i - w(\mathbf{Q}(\tau)) + N}{\|\mathbf{Q}(\tau)\|_2} \\
&\leq \frac{N - (1-\rho)w(\mathbf{Q}(\tau))}{\|\mathbf{Q}(\tau)\|_2}. \qquad (3.12)
\end{aligned}$$

The last inequality above is justified as follows. From the definition of $\rho = \rho(\boldsymbol{\lambda})$, there exist constants $\alpha_{\boldsymbol{\sigma}} \geq 0$ such that $\sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha_{\boldsymbol{\sigma}} \leq \rho$, and

$$\boldsymbol{\lambda} \leq \sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha_{\boldsymbol{\sigma}}\boldsymbol{\sigma}. \qquad (3.13)$$

Therefore,

$$\begin{aligned}
\sum_i Q_i(\tau)\lambda_i = \mathbf{Q}(\tau)\cdot\boldsymbol{\lambda} &\leq \sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha_{\boldsymbol{\sigma}}\mathbf{Q}(\tau)\cdot\boldsymbol{\sigma} \\
&\leq \sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha_{\boldsymbol{\sigma}} w(\mathbf{Q}(\tau)) \leq \rho w(\mathbf{Q}(\tau)). \qquad (3.14)
\end{aligned}$$

Let $Q_{\max}(\tau) = \max_{i=1}^N Q_i(\tau)$. Then,

$$\|\mathbf{Q}(\tau)\|_2 \leq (NQ_{\max}^2(\tau))^{\frac{1}{2}} = N^{\frac{1}{2}}Q_{\max}(\tau).$$

From Assumption 2.2.1, we have

$$w(\mathbf{Q}(\tau)) \geq Q_{\max}(\tau).$$

Therefore, the RHS of (3.12) can be upper bounded by

$$-(1-\rho)N^{-1/2} + \frac{N}{\|\mathbf{Q}(\tau)\|_2} \leq -\frac{1}{2}(1-\rho)N^{-1/2},$$

when $\|\mathbf{Q}(\tau)\|_2$ is sufficiently large.

Proof of Theorem 3.4.3: $\alpha > 1$. We now consider the case $\alpha > 1$. We wish to obtain an inequality similar to (3.12) for $L_\alpha(\mathbf{Q}(\cdot)) = \|\mathbf{Q}(\cdot)\|_{1+\alpha}$ under the MW-α policy, and we accomplish this using the second-order mean value theorem (cf. Proposition 3.3.1). Throughout this proof, we will drop the subscript $\alpha+1$ and use the notation $\|\cdot\|$ instead of $\|\cdot\|_{\alpha+1}$.

Consider the norm function

$$g(\mathbf{x}) = \|\mathbf{x}\| = (x_1^{\alpha+1} + \ldots + x_N^{\alpha+1})^{\frac{1}{\alpha+1}}.$$

The first derivative is

$$\nabla g(\mathbf{x}) = \|\mathbf{x}\|^{-\alpha}(x_1^\alpha, \ldots, x_N^\alpha) = \frac{\mathbf{x}^\alpha}{\|\mathbf{x}\|^\alpha}.$$

Let $H(\mathbf{x}) = [H_{ij}(\mathbf{x})]_{i,j=1}^N$ be the second derivative (Hessian) matrix of g. Then,

$$H_{ij}(\mathbf{x}) = \frac{\partial^2 g}{\partial x_i \partial x_j}(\mathbf{x}) = \delta_{ij}\frac{\alpha x_i^{\alpha-1}}{\|\mathbf{x}\|^\alpha} - \frac{\alpha x_i^\alpha x_j^\alpha}{\|\mathbf{x}\|^{2\alpha+1}},$$

where δ_{ij} is the Kronecker delta. By Proposition 3.3.1, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^N$, and with $\boldsymbol{\delta} = \mathbf{y} - \mathbf{x}$, there exists a $\theta \in [0,1]$ for which

$$\begin{aligned}
g(\mathbf{y}) &= g(\mathbf{x}) + \boldsymbol{\delta}^T \nabla g(\mathbf{x}) + \frac{1}{2}\boldsymbol{\delta}^T H(\mathbf{x}+\theta\boldsymbol{\delta})\boldsymbol{\delta} \\
&= g(\mathbf{x}) + \|\mathbf{x}\|^{-\alpha}\left(\sum_i \delta_i x_i^\alpha\right) + \frac{\alpha}{2}\|\mathbf{x}+\theta\boldsymbol{\delta}\|^{-\alpha}\left(\sum_i (x_i+\theta\delta_i)^{\alpha-1}\delta_i^2\right) \\
&\quad -\frac{\alpha}{2}\|\mathbf{x}+\theta\boldsymbol{\delta}\|^{-1-2\alpha}\left(\sum_{i,j}(x_i+\theta\delta_i)^\alpha(x_j+\theta\delta_j)^\alpha\delta_i\delta_j\right) \\
&= g(\mathbf{x}) + \|\mathbf{x}\|^{-\alpha}\left(\sum_i \delta_i x_i^\alpha\right) + \frac{\alpha}{2}\|\mathbf{x}+\theta\boldsymbol{\delta}\|^{-\alpha}\left(\sum_i (x_i+\theta\delta_i)^{\alpha-1}\delta_i^2\right) \\
&\quad -\frac{\alpha}{2}\|\mathbf{x}+\theta\boldsymbol{\delta}\|^{-1-2\alpha}\left(\sum_i (x_i+\theta\delta_i)^\alpha\delta_i\right)^2.
\end{aligned}$$

Using $\mathbf{x} = \mathbf{Q}(\tau)$, $\mathbf{y} = \mathbf{Q}(\tau+1)$ and $\boldsymbol{\delta}(\tau) = \mathbf{Q}(\tau+1) - \mathbf{Q}(\tau)$, we have

$$\begin{aligned} \|\mathbf{Q}(\tau+1)\| &= \|\mathbf{Q}(\tau)\| + \left[\frac{\sum_i \delta_i(\tau) Q_i^\alpha(\tau)}{\|\mathbf{Q}(\tau)\|^\alpha}\right] \\ &\quad + \frac{\alpha}{2}\left[\frac{\sum_i (Q_i(\tau) + \theta\delta_i(\tau))^{\alpha-1}\delta_i^2(\tau)}{\|\mathbf{Q}(\tau)+\theta\boldsymbol{\delta}(\tau)\|^\alpha}\right] \\ &\quad - \frac{\alpha}{2}\left[\frac{\left(\sum_i (Q_i(\tau) + \theta\delta_i(\tau))^\alpha \delta_i(\tau)\right)^2}{\|\mathbf{Q}(\tau)+\theta\boldsymbol{\delta}(\tau)\|^{1+2\alpha}}\right]. \end{aligned} \tag{3.15}$$

Therefore, using the fact that $\delta_i(\tau) \in \{-1, 0, 1\}$, we have

$$\|\mathbf{Q}(\tau+1)\| - \|\mathbf{Q}(\tau)\| \leq \left[\frac{\sum_i \delta_i(\tau) Q_i^\alpha(\tau)}{\|\mathbf{Q}(\tau)\|^\alpha}\right] + \frac{\alpha}{2}\left[\frac{\sum_i (Q_i(\tau) + \theta\delta_i(\tau))^{\alpha-1}}{\|\mathbf{Q}(\tau)+\theta\boldsymbol{\delta}(\tau)\|^\alpha}\right]. \tag{3.16}$$

We take conditional expectations of both sides, given $\mathbf{Q}(\tau)$. To bound the first term on the RHS, we use the definition of the MW-α policy, the bound (3.13) on $\boldsymbol{\lambda}$, and the argument used to establish (3.14) in the proof of Theorem 3.4.3 for $\alpha = 1$ (with $w(\mathbf{Q}(\tau))$ replaced by $w_\alpha(\mathbf{Q}(\tau))$). We obtain

$$\mathbb{E}\left[\frac{\sum_i \delta_i(\tau) Q_i^\alpha(\tau)}{\|\mathbf{Q}(\tau)\|^\alpha} \;\middle|\; \mathbf{Q}(\tau)\right] \leq -(1-\rho)\frac{w_\alpha(\mathbf{Q}(\tau))}{\|\mathbf{Q}(\tau)\|^\alpha}. \tag{3.17}$$

Note that

$$\|\mathbf{Q}(\tau)\|^\alpha \leq \left(N Q_{\max}(\tau)^{\alpha+1}\right)^{\frac{\alpha}{\alpha+1}} = N^{\frac{\alpha}{\alpha+1}} Q_{\max}^\alpha(\tau), \tag{3.18}$$

and

$$w_\alpha(\mathbf{Q}(\tau)) \geq Q_{\max}^\alpha(\tau).$$

Therefore,

$$\mathbb{E}\left[\frac{\sum_i \delta_i(\tau) Q_i^\alpha(\tau)}{\|\mathbf{Q}(\tau)\|^\alpha} \;\middle|\; \mathbf{Q}(\tau)\right] \leq -(1-\rho)N^{-\frac{\alpha}{1+\alpha}}. \tag{3.19}$$

Consider now the second term of the conditional expectation of the RHS of Inequality (3.16). Since $\alpha > 1$, and $\delta_i(\tau) \in \{-1, 0, 1\}$, the numerator of the expression

inside the bracket satisfies

$$\sum_i (Q_i(\tau) + \theta\delta_i(\tau))^{\alpha-1} \leq N\left(Q_{\max}(\tau)+1\right)^{\alpha-1},$$

and the denominator satisfies

$$\|\mathbf{Q}(\tau) + \theta\boldsymbol{\delta}(\tau)\|^{\alpha} \geq \left([Q_{\max}(\tau)-1]^{+}\right)^{\alpha},$$

where we use the notation $[c]^+ = 0 \vee c$. Thus,

$$\frac{\alpha}{2}\left[\frac{\sum_i (Q_i(\tau)+\theta\delta_i(\tau))^{\alpha-1}}{\|\mathbf{Q}(\tau)+\theta\boldsymbol{\delta}(\tau)\|^{\alpha}}\right] \leq \frac{\alpha}{2}\cdot\frac{N(Q_{\max}+1)^{\alpha-1}}{([Q_{\max}(\tau)-1]^{+})^{\alpha}}.$$

Now if $\|\mathbf{Q}(\tau)\|$ is large enough, $Q_{\max}(\tau)$ is large enough, and $\frac{\alpha}{2}\cdot\frac{N(Q_{\max}+1)^{\alpha-1}}{([Q_{\max}(\tau)-1]^{+})^{\alpha}}$ can be made arbitrarily small. Thus, the conditional expectation of the second term on the RHS of (3.16) can be made arbitrarily small for large enough $\|\mathbf{Q}(\tau)\|$. This fact, together with Inequality (3.19), implies that there exists $B > 0$ such that if $\|\mathbf{Q}(\tau)\| > B$, then

$$\mathbb{E}\left[\|\mathbf{Q}(\tau+1)\| - \|\mathbf{Q}(\tau)\| \;\middle|\; \mathbf{Q}(\tau)\right] \leq -\frac{1-\rho}{2}N^{-\frac{\alpha}{1+\alpha}}.$$

Proof of Theorem 3.4.3: $\alpha \in (0,1)$. Finally, we consider the case $\alpha < 1$. The proof in this section is similar to that for the case $\alpha > 1$. We invoke Proposition 3.3.1 to write the drift term as a sum of terms, which we bound separately. Note that to use Proposition 3.3.1, we need L_α to be twice continuously differentiable. Indeed, by Lemma 3.4.2 (i), h_α is continuously differentiable, so its antiderivative H_α is twice continuously differentiable, and so is L_α. Thus, by the second order mean value theorem, we obtain an equation similar to Equation (3.15):

$$\begin{aligned} L_\alpha(\mathbf{Q}(\tau+1)) - L_\alpha(\mathbf{Q}(\tau)) \;=\; & \left[\frac{\sum_i \delta_i(\tau)h_\alpha(Q_i(\tau))}{L_\alpha^{\alpha}(\mathbf{Q}(\tau))}\right] + \frac{1}{2}\left[\frac{\sum_i h'_\alpha(Q_i(\tau)+\theta\delta_i(\tau))\delta_i^2(\tau)}{L_\alpha^{\alpha}(\mathbf{Q}(\tau)+\theta\boldsymbol{\delta}(\tau))}\right] \\ & -\frac{\alpha}{2}\left[\frac{(\sum_i \delta_i(\tau)h_\alpha(Q_i(\tau)+\theta\delta_i(\tau)))^2}{L_\alpha^{2\alpha+1}(\mathbf{Q}(\tau)+\theta\boldsymbol{\delta}(\tau))}\right]. \end{aligned} \tag{3.20}$$

Again, using the fact $\delta_i(\tau) \in \{-1, 0, 1\}$,

$$L_\alpha(\mathbf{Q}(\tau+1)) - L_\alpha(\mathbf{Q}(\tau)) \leq T_1 + T_2,$$

where

$$T_1 = \frac{\sum_i \delta_i(\tau) h_\alpha(Q_i(\tau))}{L_\alpha^\alpha(\mathbf{Q}(\tau))},$$

and

$$T_2 = \frac{1}{2}\left[\frac{\sum_i h'_\alpha(Q_i(\tau) + \theta\delta_i(\tau))}{L_\alpha^\alpha(\mathbf{Q}(\tau) + \theta\boldsymbol{\delta}(\tau))}\right].$$

Let us consider T_2 first. For $\alpha \in (0,1)$, by Lemma 3.4.2 (iv), $h'_\alpha(r) \leq 2$ for all $r \geq 0$. Thus

$$T_2 \leq \frac{1}{2}\left[\frac{2N}{L_\alpha^\alpha(\mathbf{Q}(\tau) + \theta\boldsymbol{\delta}(\tau))}\right] = \frac{N}{L_\alpha^\alpha(\mathbf{Q}(\tau) + \theta\boldsymbol{\delta}(\tau))}.$$

which becomes arbitrarily small when $L_\alpha(\mathbf{Q}(\tau))$ is large enough.

We now consider T_1. Since $h_\alpha(r) \leq r^\alpha + 1$ for all $r \geq 0$ (cf. Lemma 3.4.2 (iii)), and $\delta_i(\tau) \in \{-1, 0, 1\}$,

$$T_1 \leq \frac{\sum_i \delta_i(\tau) Q_i^\alpha(\tau)}{L_\alpha^\alpha(\mathbf{Q}(\tau))} + \frac{N}{L_\alpha^\alpha(\mathbf{Q}(\tau))}.$$

When we take the conditional expectation, an argument similar to the one for the case $\alpha > 1$ yields

$$\mathbb{E}\left[\frac{\sum_i \delta_i(\tau) Q_i^\alpha(\tau)}{L_\alpha^\alpha(\mathbf{Q}(\tau))} \;\middle|\; \mathbf{Q}(\tau)\right] \leq -(1-\rho)\frac{w_\alpha(\mathbf{Q}(\tau))}{L_\alpha^\alpha(\mathbf{Q}(\tau))}. \tag{3.21}$$

Again, as before, $w_\alpha(\mathbf{Q}(\tau)) \geq Q_{\max}^\alpha(\tau)$. For the denominator, by Lemma 3.4.2 (iii'), for any $r \geq 0$, we have $(\alpha+1)H_\alpha(r) \leq r^{\alpha+1} + 2$. Thus

$$\begin{aligned} L_\alpha(\mathbf{Q}(\tau)) &\leq \left[\sum_i (Q_i(\tau)+2)^{\alpha+1}\right]^{\frac{1}{\alpha+1}} \\ &\leq \left(N(Q_{\max}(\tau)+2)^{\alpha+1}\right)^{\frac{1}{\alpha+1}} \\ &= N^{\frac{1}{\alpha+1}}(Q_{\max}(\tau)+2). \end{aligned}$$

Therefore,

$$\mathbb{E}\left[\frac{\sum_i \delta_i(\tau) Q_i^\alpha(\tau)}{L_\alpha^\alpha \mathbf{Q}(\tau)} \;\middle|\; \mathbf{Q}(\tau)\right] \leq -(1-\rho)N^{-\frac{\alpha}{\alpha+1}} \frac{Q_{\max}^\alpha(\tau)}{(Q_{\max}+2)^\alpha}.$$

If $Q_{\max}(\tau)$ is large enough, we can further upper bound the RHS by, say, $-\frac{3}{4}(1-\rho)N^{-\frac{\alpha}{\alpha+1}}$.

Putting everything together, we have

$$\begin{aligned}
&\mathbb{E}\left[L_\alpha(\mathbf{Q}(\tau+1)) - L_\alpha(\mathbf{Q}(\tau)) \;\middle|\; \mathbf{Q}(\tau)\right] \\
&\leq -\frac{3}{4}(1-\rho)N^{-\frac{\alpha}{1+\alpha}} + \frac{N}{L_\alpha^\alpha(\mathbf{Q}(\tau))} + \mathbb{E}[T_2 \mid \mathbf{Q}(\tau)], \qquad (3.22)
\end{aligned}$$

if $Q_{\max}(\tau)$ is large enough. As before, if $L_\alpha(\mathbf{Q}(\tau))$ is large enough, then $Q_{\max}(\tau)$ is large enough, and T_2 and $\frac{N}{L_\alpha^\alpha(\mathbf{Q}(\tau))}$ can be made arbitrarily small. Thus, there exists $B > 0$ such that if $L_\alpha(\mathbf{Q}(\tau)) > B$, then

$$\mathbb{E}\left[L_\alpha(\mathbf{Q}(\tau+1)) - L_\alpha(\mathbf{Q}(\tau)) \;\middle|\; \mathbf{Q}(\tau)\right] \leq -\frac{1}{2}(1-\rho)N^{-\frac{\alpha}{1+\alpha}}.$$

3.4.2 Exponential Bound under MW-α

In this section we derive an exponential upper bound on the tail probability of the stationary queue-size distribution, under the MW-α policy.

The proof of Theorem 3.2.2 relies on the drift inequality obtained in Theorem 3.4.3, and the following theorem, a modification of Theorem 1 from [4].

Theorem 3.4.4 *Let $\mathbf{X}(\cdot)$ be an irreducible and aperiodic discrete-time Markov chain with a countable state space $\mathscr{X}$. Suppose that there exists a Lyapunov function $f : \mathscr{X} \to \mathbb{R}_+$ with the following properties:*

*(a) f has **bounded increments**: there exists $\xi > 0$ such that for all τ, we have*

$$|f(\mathbf{X}(\tau+1)) - f(\mathbf{X}(\tau))| \leq \xi, \;\; \textit{almost surely} \; ;$$

(b) ***Negative drift****: there exist $B > 0$ and $\gamma > 0$ such that whenever $f(\mathbf{X}(\tau)) > B$,*

$$\mathbb{E}[f(\mathbf{X}(\tau+1)) - f(\mathbf{X}(\tau)) \mid \mathbf{X}(\tau)] \leq -\gamma.$$

Then, a stationary probability distribution π exists, and we have an exponential upper bound on the tail probability of f under π: for any $\ell \in \mathbb{Z}_+$,

$$\mathbb{P}_\pi(f(\mathbf{X}) > B + 2\xi\ell) \leq \left(\frac{\xi}{\xi+\gamma}\right)^{\ell+1}. \tag{3.23}$$

In particular, in steady state, all moments of f are finite, i.e., for every $k \in \mathbb{N}$,

$$\mathbb{E}_\pi[f^k(\mathbf{X})] < \infty.$$

Theorem 3.4.4 is identical to Theorem 1 in [4] except that [4] imposed the additional condition $\mathbb{E}_\pi[f(\mathbf{X})] < \infty$. However, the latter condition is redundant. Indeed, using Foster-Lyapunov criteria (see [18], for example), conditions (a) and (b) in Theorem 3.4.4 imply that the Markov chain $\mathbf{X}$ has a unique stationary distribution π. Furthermore, Theorem 2.3 in [27] establishes that under conditions (a) and (b), all moments of $f(\mathbf{X})$ are finite in steady state. We note that Theorem 2.3 in [27] and Theorem 1 of [4] provide the same qualitative information (exponential tail bounds for $f(\mathbf{X})$). However, [4] contains the tighter and explicit bound (3.23), which we use here to prove Theorem 3.2.2.

Proof of Theorem 3.2.2. When $\alpha \geq 1$, the proof of Theorem 3.2.2 follows immediately from Theorem 3.4.4 by noticing that Theorem 3.4.3 provides the desired drift inequality, and the maximal change in $\|\mathbf{Q}(\tau)\|_{1+\alpha}$ in one time step is at most $\nu_{\max} = N^{\frac{1}{1+\alpha}}$, because each queue can receive at most one arrival and have at most one departure per time step. The proof for the case when $\alpha \in (0,1)$ is entirely parallel, and we do not reproduce it here.

3.5 Tightness of the Exponential Upper Bound

Given the exponential upper bound, it is natural to ask how tight the bound is. Here we compare the exponential upper bound in Theorem 3.2.2 with a universal lower bound for any online scheduling policy. To be able to evaluate a useful lower bound explicitly, we consider the case of an input-queued switch. As discussed in Section 2.2.1, Chapter 2, in an $n \times n$ input-queued switch, there are $N = n^2$ queues. Recall that if we let $\boldsymbol{\lambda} = [\lambda_{ij}]_{i,j=1}^n$ be the arrival rate matrix, then the load $\rho = \rho(\boldsymbol{\lambda})$ is given by

$$\rho(\boldsymbol{\lambda}) = \max_{1 \le k,\ell \le n} \left\{ \sum_{m=1}^n \lambda_{k,m}, \sum_{m=1}^n \lambda_{m,\ell} \right\}.$$

Upper Bound for Input-Queued Switches. We consider an $n \times n$ input-queued switch operating under a MW-α policy, where $\alpha > 0$. Let the load be $\rho \in (0,1)$. Then the Markov chain $\mathbf{Q}(\cdot)$ is positive recurrent, and a unique stationary distribution $\boldsymbol{\pi}$ exists for $\mathbf{Q}(\cdot)$.

The quantity of interest is $\frac{1}{K}\mathbb{P}_{\boldsymbol{\pi}}\left(\|\mathbf{Q}\|_1 \ge K\right)$, when K is large. Here we are interested in

$$\limsup_{K\to\infty} \frac{1}{K}\mathbb{P}_{\boldsymbol{\pi}}\left(\|\mathbf{Q}\|_1 \ge K\right).$$

Note that by Theorem 3.2.2, when $\rho \to 1$, and using the relation $\log(1+r) \approx r$ for small $r > 0$, we have

$$\limsup_{K\to\infty} \frac{1}{K}\mathbb{P}_{\boldsymbol{\pi}}\left(\|\mathbf{Q}\|_{\alpha+1} \ge K\right) \lesssim -\frac{1-\rho}{100} N^{-1-\frac{1}{1+\alpha}} = -\frac{1-\rho}{100} n^{-2-\frac{2}{1+\alpha}},$$

using the relation $N = n^2$. By Jensen's inequality, and using the convexity of the function $x \mapsto x^{1+\alpha}$ for $x \ge 0$, we have that for any $\mathbf{x} \in \mathbb{R}^N$ and any $\alpha > 0$,

$$N^{\frac{\alpha}{\alpha+1}} \|\mathbf{x}\|_{\alpha+1} \ge \|\mathbf{x}\|_1.$$

Thus for any $K > 0$,

$$\frac{1}{K}\mathbb{P}_{\boldsymbol{\pi}}\left(\|\mathbf{Q}\|_1 \ge K\right) \le \frac{1}{K}\mathbb{P}_{\boldsymbol{\pi}}\left(\|\mathbf{Q}\|_{\alpha+1} \ge \frac{K}{N^{\alpha/(\alpha+1)}}\right)$$

$$= \frac{1}{N^{\alpha/(\alpha+1)}} \cdot \frac{N^{\alpha/(\alpha+1)}}{K} \mathbb{P}_{\boldsymbol{\pi}} \left(\|\mathbf{Q}\|_{\alpha+1} \geq \frac{K}{N^{\alpha/(\alpha+1)}} \right),$$

and so

$$\begin{aligned} \limsup_{K\to\infty} \frac{1}{K} \mathbb{P}_{\boldsymbol{\pi}} \left(\|\mathbf{Q}\|_1 \geq K \right) &\leq \frac{1}{N^{\alpha/(\alpha+1)}} \limsup_{K'\to\infty} \frac{1}{K'} \mathbb{P}_{\boldsymbol{\pi}} \left(\|\mathbf{Q}\|_{\alpha+1} \geq K' \right) \\ &\lesssim -\frac{1-\rho}{100} n^{-2-\frac{2}{1+\alpha}} N^{-\alpha/(\alpha+1)} = -\frac{1-\rho}{100} n^{-4}. \end{aligned}$$

We would like to note that in the Appendix of Shah et al. [53], a tighter upper bound of $-\frac{1-\rho}{100} n^{-3}$ can be obtained, through a similar but more refined analysis.

Lower Bound for Input-Queued Switches. We now present a universal exponential lower bound for any online policy in input-queued switches. Here we suppose that the arrival rate matrix is uniform, i.e., for all $i, j \in \{1, 2, \ldots, n\}$, $\lambda_{ij} = \rho/n$, where $\rho \in (0, 1)$. The load associated with this arrival rate matrix is then ρ.

We concentrate on a single output port, and the steady-state total number of packets associated with this output port. In each time slot, the total number of arrivals to this output port is a binomial random variable with parameters n and ρ/n, and at most one packet can depart the output port. Thus, the total number of packets associated with this output port in steady state stochastically dominates the same quantity in a Bin/D/1 queue (there is potentially more service provided for the Bin/D/1 queue).

We now consider the steady-state queue length Q_{Bin} of this Bin/D/1 queue, and its large-deviations exponent

$$\lim_{K\to\infty} \frac{1}{K} \mathbb{P} \left(Q_{\text{Bin}} \geq K \right).$$

By Theorem 1.4 of [23], the limit is well-defined, and is given by $-\theta^*$, where θ^* is defined by

$$\theta^* = \sup\{\theta > 0 : M(\theta) < \theta\},$$

where $M(\theta) = \log \mathbb{E}[e^{\theta X}]$ is the log-moment generating function of a random variable

X that has a binomial distribution with parameters n and ρ/n. It is well-known that

$$M(\theta) = n \log \left(1 - \frac{\rho}{n} + \frac{\rho}{n} e^{\theta}\right).$$

Thus, θ^* must satisfy

$$\theta^* = n \log \left(1 - \frac{\rho}{n} + \frac{\rho}{n} e^{\theta^*}\right).$$

While a closed-form solution of θ^* seems impossible, we can look for an approximation. We are interested in comparing the bounds for large n and ρ near 1, and we will develop a good approximation in that regime. We expect the optimizing value θ^* to be small, in which case $e^{\theta^*} \approx 1 + \theta^* + \frac{(\theta^*)^2}{2}$, and

$$\log \left(1 - \frac{\rho}{n} + \frac{\rho}{n} e^{\theta^*}\right) \approx \frac{\rho}{n}(e^{\theta^*} - 1) \approx \frac{\rho}{n}(\theta^* + \frac{(\theta^*)^2}{2}).$$

Thus, we expect that

$$n \cdot \frac{\rho}{n} \left(\theta^* + \frac{(\theta^*)^2}{2}\right) \approx \theta^*,$$

which leads to an optimal solution $\theta^* \approx 2(1 - \rho)$. This shows that, in particular,

$$\liminf_{K \to \infty} \frac{1}{K} \mathbb{P}\left(Q_{\text{Bin}} \geq K\right) \gtrsim -2(1 - \rho).$$

Since Q_{Bin} is stochastically dominated by the total number of packets at an output port in steady state, Q_{Bin} is also stochastically dominated by $\|\mathbf{Q}\|_1$, under any online scheduling policy. This shows that

$$\liminf_{K \to \infty} \frac{1}{K} \mathbb{P}\left(\|\mathbf{Q}(\infty)\|_1 \geq K\right) \gtrsim -2(1 - \rho),$$

under any online scheduling policy, in an $n \times n$ input-queued switch, and when n is large and ρ is close to 1.

Comparison. For any $\alpha > 0$, consider the upper bound $-\frac{1-\rho}{100} n^{-4}$ (or $-\frac{1-\rho}{100} n^{-3}$), and the lower bound $-2(1 - \rho)$. Ignoring universal constants, the ratio between the bounds is precisely n^4 (or n^3). From this, we see that the dependence of our upper

bound exponent on the load ρ is tight, when the system is heavily loaded. However, the dependence on the number $N = n^2$ of queues is not. An immediate question is whether the dependence on N can be made tighter. First, are the exponential upper bounds obtained in this chapter tight, in that the tail exponent under MW-α is $O(-(1-\rho)n^{-3})$? We expect the answer to be negative, and we suspect that MW-α produces a $O\big(-(1-\rho)\big)$ tail exponent, for each $\alpha > 0$. Second, can we at least design a policy that provably achieves a tight tail exponent of order $O(-(1-\rho))$? In Chapter 5, we will see that such a policy does exist, with a tail exponent $\approx -2(1-\rho)/\rho$.

Chapter 4

Performance of α-Fair Policies in BN

In this chapter, we establish various qualitative performance bounds for so-called α-fair policies in bandwidth-sharing networks. The structure of this chapter is similar to Chapter 3. We first define α-fair policies in Section 4.1, followed by some preliminaries in Section 4.2. We then state our main results in Section 4.3. In Section 4.4, we present a transient analysis of the α-fair policies, for $\alpha \geq 1$. Note that Lemma 3.3.2 is used again to prove a maximal inequality under the α-fair policy, for $\alpha \geq 1$ (Theorem 4.3.1). We then apply the maximal inequality to establish the full state space collapse property for $\alpha \geq 1$ (Theorem 4.4.9). In Section 4.5, we present an exponential upper bound on the tail probability of the steady-state distribution under the α-fair policy, for $\alpha \in (0, \infty)$. We start by establishing a drift inequality, Theorem 4.5.3, for a suitably defined "normed" Lyapunov function (Definition 4.5.1). This drift inequality is crucial for proving the exponential upper bound, Theorem 4.3.2.

Section 4.6 contains an application of Theorem 4.3.2. Building upon previous work by Kang et al. [32], we use the exponential upper bound to establish the validity of the diffusion approximation in steady state, Theorem 4.6.6, for a bandwidth-sharing network under a proportionally fair policy ($\alpha = 1$). This leads to an elegant product-form description of the limit of the diffusion-scaled steady-state distributions. In Section 4.7, we use this product form to perform a formal calculation of the steady-

state performance of proportional fairness for input-queued switches, and conjecture its optimality in this model (Conjecture 4.7.1). We conclude the chapter with some discussion in Section 4.8.

The prerequisite for reading this chapter is the description of the bandwidth-sharing network model in Section 2.3, Chapter 2.

4.1 The α-Fair Bandwidth-Sharing Policy

A bandwidth sharing policy has to allocate rates to flows so that capacity constraints are satisfied at each time instance. Here we discuss the popular α-fair bandwidth-sharing policy, where $\alpha > 0$. At any time, the bandwidth allocation depends on the current number of flows $\mathbf{m} = (m_i)_{i\in\mathcal{I}}$. Let ϕ_i be the *total* bandwidth allocated to route i under the α-fair policy: each flow of type i gets rate ϕ_i/m_i if $m_i > 0$, and $\phi_i = 0$ if $m_i = 0$. Under an α-fair policy, the bandwidth vector $\boldsymbol{\phi}(\mathbf{m}) = (\phi_i(\mathbf{m}))_{i\in\mathcal{I}}$ is determined as follows.

If $\mathbf{m} = \mathbf{0}$, then $\boldsymbol{\phi} = \mathbf{0}$. If $\mathbf{m} \neq \mathbf{0}$, let $\mathcal{I}_+(\mathbf{m}) = \{i \in \mathcal{I} : m_i > 0\}$. For $i \notin \mathcal{I}_+(\mathbf{m})$, set $\phi_i(\mathbf{m}) = 0$. Let $\boldsymbol{\phi}_+(\mathbf{m}) = (\phi_i(\mathbf{m}))_{i\in\mathcal{I}_+(\mathbf{m})}$. Then, $\boldsymbol{\phi}_+(\mathbf{m})$ is the *unique* maximizer in the optimization problem

$$\text{maximize} \quad G_n(\boldsymbol{\phi}_+) \qquad \text{over} \quad \boldsymbol{\phi} \in \mathbb{R}_+^N \tag{4.1}$$

$$\text{subject to} \quad \sum_{i\in\mathcal{I}_+(\mathbf{m})} R_{ji}\phi_i \leq C_j, \quad \forall\, j \in \mathcal{J}, \tag{4.2}$$

where

$$G_{\mathbf{m}}(\boldsymbol{\phi}_+) = \begin{cases} \displaystyle\sum_{i\in\mathcal{I}_+(\mathbf{m})} \kappa_i m_i^{\alpha} \frac{\phi_i^{1-\alpha}}{1-\alpha}, & \text{if } \alpha \in (0,\infty)\backslash\{1\}, \\ \displaystyle\sum_{i\in\mathcal{I}_+(\mathbf{m})} \kappa_i m_i \log \phi_i, & \text{if } \alpha = 1. \end{cases}$$

Here, for each $i \in \mathcal{I}$, κ_i is a positive weight assigned to route i.

Flow Dynamics. Recall from Section 2.3 that flows arrive according to a Poisson process, and each arriving flow brings an amount of work that is exponentially

distributed. The flow dynamics are described by the evolution of the flow vector $\mathbf{M}(t) = (M_i(t))_{i \in \mathcal{I}}$, a Markov process with infinitesimal transition rate matrix $\mathbf{q}$ given by

$$q(\mathbf{n}, \mathbf{n}+\mathbf{m}) = \begin{cases} \nu_i, & \text{if } \mathbf{m} = \mathbf{e}_i, \\ \mu_i \Lambda_i(\mathbf{n}), & \text{if } \mathbf{m} = -\mathbf{e}_i, \text{ and } n_i \geq 1, \\ 0, & \text{otherwise,} \end{cases} \tag{4.3}$$

where for each i, $\nu_i > 0$ and $\mu_i > 0$ are the arrival and service rates defined in Section 2.3, and $\mathbf{e}_i$ is the i-th unit vector.

4.2 Preliminaries

A Note on Our Use of Constants. Our results and proofs involve various constants; some are absolute constants, some depend only on the structure of the network, and some depend (smoothly) on the traffic parameters (the arrival and service rates). It is convenient to distinguish between the different types of constants, and we define here the terminology that we will be using.

The term *absolute constant* will be used to refer to a quantity that does not depend on any of the model parameters. The term *network-dependent constant* will be used to refer to quantities that are completely determined by the structure of the underlying network and policy, namely, the incidence matrix $\mathbf{R}$, the capacity vector $\mathbf{C}$, the weight vector $\boldsymbol{\kappa}$, and the policy parameter α.

Our analysis also involves certain quantities that depend on the traffic parameters, namely, the arrival and service parameters $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. These quantities are often given by complicated expressions that would be inconvenient to carry through the various arguments. It turns out that the only property of such quantities that is relevant to our purposes is the fact they change continuously as $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ vary over the open positive orthant. (This still allows these quantities to be undefined or discontinuous on the boundary of the positive orthant.) We abstract this property by introducing, in the definition that follows, the concept of a *(positive) load-dependent constant.*

Definition 4.2.1 *Consider a family of bandwidth-sharing networks with common parameters* $(\mathbf{R}, \mathbf{C}, \boldsymbol{\kappa}, \alpha)$*, but varying traffic parameters* $(\boldsymbol{\mu}, \boldsymbol{\nu})$*. A quantity* K *will be called a* (positive) load-dependent constant *if for networks in that family it is determined by a relation of the form* $K = f(\boldsymbol{\mu}, \boldsymbol{\nu})$*, where* $f : R_p^N \times R_p^N \to R_p$ *is a* continuous *function on the open positive orthant* $R_p^N \times R_p^N$*.*

A key property of a load-dependent constant, which will be used in some of the subsequent proofs, is that it is by definition positive and furthermore (because of continuity), bounded above and below by positive network-dependent constants if we restrict $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ to a compact subset of the open positive orthant. A natural example of a load-dependent constant is the load factor $\lambda_i = \nu_i/\mu_i$. (Note that this quantity diverges as $\mu_i \to 0$.) We also define the *gap* of a underloaded bandwidth-sharing network.

Definition 4.2.2 *Consider a family of bandwidth-sharing networks with common parameters* $(\mathbf{R}, \mathbf{C}, \boldsymbol{\kappa}, \alpha)$ *and with varying traffic parameters* $(\boldsymbol{\mu}, \boldsymbol{\nu})$ *that satisfy* $\mathbf{R}\boldsymbol{\lambda} < \mathbf{C}$*. The* gap *of a network with traffic parameters* $(\boldsymbol{\mu}, \boldsymbol{\nu})$ *in the family, denoted by* $\varepsilon(\boldsymbol{\lambda})$*, is defined by*

$$\varepsilon(\boldsymbol{\lambda}) \triangleq \sup\{\tilde{\varepsilon} > 0 : (1 + \tilde{\varepsilon})\mathbf{R}\boldsymbol{\lambda} \leq \mathbf{C}\}.$$

We sometimes write ε for $\varepsilon(\boldsymbol{\lambda})$ when there is no ambiguity. Note also that $\varepsilon(\boldsymbol{\lambda})$ plays the same role as the term $1 - \rho$ in a queueing system with load ρ.

Uniformization. Uniformization is a well-known device which allows us to study a continuous-time Markov process by considering an associated discrete-time Markov chain with the same stationary distribution. We provide here some details, and the notation that we will be using.

Recall that the Markov process $\mathbf{M}(\cdot)$ of interest has dynamics given by (4.3). Let $\Xi(\mathbf{m}) = \sum_{\tilde{\mathbf{m}}} q(\mathbf{m}, \tilde{\mathbf{m}})$ be the aggregate transition rate at state $\mathbf{m}$. The *embedded jump chain* of $\mathbf{M}(\cdot)$ is a discrete-time Markov chain with the same state space $\mathbb{Z}_+^N$,

and with transition probability matrix $\mathbf{P}$ given by

$$P(\mathbf{m}, \tilde{\mathbf{m}}) = \frac{q(\mathbf{m}, \tilde{\mathbf{m}})}{\Xi(\mathbf{m})}.$$

The so-called *uniformized Markov chain* is an alternative, more convenient, discrete-time Markov chain, denoted $\left(\tilde{\mathbf{M}}(\tau)\right)_{\tau \in \mathbb{Z}_+}$, to be defined shortly.

We first introduce some more notation. Consider the aggregate transition rates $\Xi(\mathbf{m}) = \sum_{\tilde{\mathbf{m}}} q(\mathbf{m}, \tilde{\mathbf{m}})$. Since every route uses at least one resource, we have $\phi_i(\mathbf{m}) \leq \max_{j \in \mathcal{J}} C_j$, for all $i \in \mathcal{I}$. Then, by (4.3), we have

$$\Xi(\mathbf{m}) = \sum_{\tilde{\mathbf{m}}} q(\mathbf{m}, \tilde{\mathbf{m}}) \leq \sum_{i \in \mathcal{I}} (\nu_i + \mu_i \phi_i(\mathbf{m})) \leq \sum_{i \in \mathcal{I}} \left(\nu_i + \mu_i \max_{j \in \mathcal{J}} C_j \right).$$

We define $\Xi \triangleq \sum_{i \in \mathcal{I}} (\nu_i + \mu_i \max_{j \in \mathcal{J}} C_j)$, and modify the rates of self-transitions (which were zero in the original model) to

$$q(\mathbf{m}, \mathbf{m}) := \Xi - \Xi(\mathbf{m}). \tag{4.4}$$

Note that Ξ is a positive load-dependent constant. We define a transition probability matrix $\tilde{\mathbf{P}}$ by

$$\tilde{P}(\mathbf{m}, \tilde{\mathbf{m}}) \triangleq \frac{q(\mathbf{m}, \tilde{\mathbf{m}})}{\Xi}.$$

Definition 4.2.3 *The uniformized Markov chain* $\left(\tilde{\mathbf{M}}(\tau)\right)_{\tau \in \mathbb{Z}_+}$ *associated with the Markov process* $\mathbf{M}(\cdot)$ *is a discrete-time Markov chain with the same state space* $\mathbb{Z}_+^N$, *and with transition matrix* $\tilde{\mathbf{P}}$ *defined as above.*

As remarked earlier, the Markov process $\mathbf{M}(\cdot)$ that describes a bandwidth-sharing network operating under an α-fair policy is positive recurrent, as long as the system is underloaded, i.e., if $\mathbf{R}\boldsymbol{\lambda} < \mathbf{C}$. It is not hard to verify that $\mathbf{M}(\cdot)$ is also irreducible. Therefore, the Markov process $\mathbf{M}(\cdot)$ has a unique stationary distribution. The chain $\tilde{\mathbf{M}}(\cdot)$ is also positive recurrent and irreducible, because $\mathbf{M}(\cdot)$ is, and by suitably increasing Ξ if necessary, it can be made aperiodic. Thus $\tilde{\mathbf{M}}(\cdot)$ has a unique stationary distribution as well. A crucial property of the uniformized chain $\tilde{\mathbf{M}}(\cdot)$ is that this

unique stationary distribution is the same as that of the original Markov process $\mathbf{M}(\cdot)$; see, e.g., [19].

4.3 Summary of Results

In this section, we summarize our main results for both the transient and the steady-state regime. The proofs are given in subsequent sections.

4.3.1 Transient Regime

Here we provide a simple inequality on the maximal excursion of the number of flows over a finite time interval, under an α-fair policy with $\alpha \geq 1$.

Theorem 4.3.1 *Consider a bandwidth-sharing network operating under an α-fair policy with $\alpha \geq 1$, and assume that $\mathbf{R}\boldsymbol{\lambda} < \mathbf{C}$. Suppose that $\mathbf{M}(0) = \mathbf{0}$. Let $N^*(T) = \sup_{t\in[0,T],i\in\mathcal{I}} M_i(t)$, and let ε be the gap. Then, for any $b > 0$,*

$$\mathbb{P}\left(N^*(T) \geq b\right) \leq \frac{KT}{\varepsilon^{\alpha-1}b^{\alpha+1}}, \tag{4.5}$$

for some positive load-dependent constant K.

As an important application, in Section 4.4.3, we will use Theorem 4.3.1 to prove a full state space collapse result, when $\alpha \geq 1$. (As discussed in the introduction, this property is stronger than multiplicative state space collapse.) The precise statement can be found in Theorem 4.4.9.

4.3.2 Stationary Regime

As noted earlier, the Markov process $\mathbf{M}(\cdot)$ has a unique stationary distribution, which we will denote by $\boldsymbol{\pi}$. We use $\mathbb{E}_{\boldsymbol{\pi}}$ and $\mathbb{P}_{\boldsymbol{\pi}}$ to denote expectations and probabilities under $\boldsymbol{\pi}$.

Exponential Bound on Tail Probabilities. For an α-fair policy, and for any $\alpha \in (0, \infty)$, we obtain an explicit exponential upper bound on the tail probabilities for the number of flows, in steady state. This will be used to establish an "interchange of limits" result in Section 4.6. See Theorem 4.6.6 for more details.

Theorem 4.3.2 *Consider a bandwidth-sharing network operating under an α-fair policy with $\alpha > 0$, and assume that $\mathbf{R}\boldsymbol{\lambda} < \mathbf{C}$. Let ε be the gap. There exist positive constants B, K, and ξ such that for all $\ell \in \mathbb{Z}_+$:*

$$\mathbb{P}_\pi \left(\|\mathbf{M}\|_\infty \geq B + 2\xi\ell \right) \leq \left(\frac{\xi}{\xi + \varepsilon K} \right)^{\ell+1}. \tag{4.6}$$

Here ξ and K are load-dependent constants, and B takes the form K'/ε when $\alpha \geq 1$, and $K'/\min\{\varepsilon^{1/\alpha}, \varepsilon\}$ when $\alpha \in (0,1)$, with K' being a load-dependent constant. In particular, all moments of $\|\mathbf{M}\|_\infty$ are finite under the stationary distribution $\boldsymbol{\pi}$, i.e., $\mathbb{E}_{\boldsymbol{\pi}}[\|\mathbf{M}\|_\infty^k] < \infty$ for every $k \in \mathbb{N}$.

Here we note that Theorem 4.3.2 implies the following. The system load $\rho(\boldsymbol{\lambda})$ defined in Section 2.3, satisfies $\rho(\boldsymbol{\lambda}) \triangleq \frac{1}{1+\varepsilon(\boldsymbol{\lambda})}$. If $\varepsilon = \varepsilon(\boldsymbol{\lambda})$ is small, i.e., if the system approaches criticality, then $\rho(\boldsymbol{\lambda}) \approx 1 - \varepsilon(\boldsymbol{\lambda})$. In this case, an immediate consequence of the bound (4.6) is that

$$\begin{aligned} \limsup_{\gamma \to \infty} \frac{1}{\gamma} \log \mathbb{P}_\pi(\|\mathbf{M}\|_\infty \geq \gamma) &\lesssim \frac{1}{2\xi} \log \left(\frac{\xi}{\xi + \varepsilon K} \right) \\ &\approx -\frac{K\varepsilon}{2\xi^2} \approx -\frac{K}{2\xi^2}(1 - \rho(\boldsymbol{\lambda})). \end{aligned}$$

Note that $\frac{K}{2\xi^2}$ is a load-dependent constant. Thus Theorem 4.3.2 shows that the large-deviations exponent of the steady-state number of flows is upper bounded by $-(1 - \rho(\boldsymbol{\lambda}))$, up to a multiplicative load-dependent constant.

Interchange of Limits ($\alpha = 1$). As discussed in the introduction, when $\alpha = 1$, Theorem 4.3.2 leads to the tightness (Lemma 4.6.7) of the steady-state distributions of the model under diffusion scaling. This in turn leads to Theorem 4.6.6 and Corollary 4.6.11, on the validity of the diffusion approximation in steady state. As the

statements of these results require a significant amount of preliminary notation and background (which is introduced in Section 4.6), we give here an informal statement.

INTERCHANGE OF LIMITS THEOREM (*informal statement*): *Consider a sequence of flow-level networks operating under the proportionally fair policy. Let* $\mathbf{M}^r(\cdot)$ *be the flow-vector Markov process associated with the* r*th network, let* ε^r *be the corresponding gap, and let* $\hat{\boldsymbol{\pi}}^r$ *be the stationary distribution of* $\varepsilon^r \mathbf{M}^r(\cdot)$. *As* $\varepsilon^r \to 0$, *and under certain technical conditions,* $\hat{\boldsymbol{\pi}}^r$ *converges weakly to the stationary distribution of an associated limiting process.*

4.4 Transient Analysis ($\alpha \geq 1$)

In this section, we present a transient analysis of the α-fair policies with $\alpha \geq 1$. First we present a general maximal lemma, which we then specialize to our model. In particular, we prove a refined drift inequality for the Lyapunov function given by

$$F_\alpha(\mathbf{m}) = \frac{1}{\alpha+1} \sum_{i \in \mathcal{I}} \nu_i \kappa_i \mu_i^{\alpha-1} \left(\frac{m_i}{\nu_i} \right)^{\alpha+1}. \tag{4.7}$$

This Lyapunov function and associated drift inequalities have played an important role in establishing positive recurrence (cf. [6], [15], [36]) and multiplicative state space collapse (cf. [32]) for α-fair policies. We combine our drift inequality with the maximal lemma to obtain a maximal inequality for bandwidth-sharing networks. We then apply the maximal inequality to prove full state space collapse when $\alpha \geq 1$.

4.4.1 The Key Lemma

The analysis here depends on the same maximal lemma, Lemma 3.3.2 as in Section 3.3.1, Chapter 3, which we repeat here for easy reference.

Lemma 4.4.1 *Let* $(\mathscr{F}_n)_{n \in \mathbb{Z}_+}$ *be a filtration on a probability space. Let* $(X_n)_{n \in \mathbb{Z}_+}$ *be a nonnegative* $\mathscr{F}_n$*-adapted stochastic process that satisfies*

$$\mathbb{E}[X_{n+1} \mid \mathscr{F}_n] \leq X_n + B_n \tag{4.8}$$

where the B_n are nonnegative random variables (not necessarily $\mathscr{F}_n$-adapted) with finite means. Let $X_n^* = \max\{X_0, \ldots, X_n\}$ and suppose that $X_0 = 0$. Then, for any $a > 0$ and any $T \in \mathbb{Z}_+$,

$$\mathbb{P}(X_T^* \geq a) \leq \frac{\sum_{n=0}^{T-1} \mathbb{E}[B_n]}{a}.$$

Since we are dealing with continuous-time Markov processes, the following corollary of Lemma 4.4.1 will be useful for our analysis.

Corollary 4.4.2 *Let $(\mathscr{F}_t)_{t\geq 0}$ be a filtration on a probability space. Let Z_t be a nonnegative, right-continuous $\mathscr{F}_t$-adapted stochastic process that satisfies*

$$\mathbb{E}\left[Z_{s+t}|\mathscr{F}_s\right] \leq Z_s + Bt,$$

for all $s, t \geq 0$, where B is a nonnegative constant. Assume that $Z_0 \equiv 0$. Denote $Z_T^* \triangleq \sup_{0\leq t\leq T} Z_t$ (which can possibly be infinite). Then, for any $a > 0$, and for any $T \geq 0$,

$$\mathbb{P}(Z_T^* \geq a) \leq \frac{BT}{a}.$$

Proof. The proof is fairly standard. We fix $T \geq 0$ and $a > 0$. Since Z_t is right-continuous, $Z_T^* = \sup_{t\in[0,T]} Z_t = \sup_{t\in([0,T]\cap\mathbb{Q})\cup\{T\}} Z_t$. Consider an increasing sequence of finite sets I_n so that $\cup_{n=1}^{\infty} I_n = ([0,T] \cap \mathbb{Q}) \cup \{T\}$, and $0, T \in I_n$ for all n. Define $Z_T^{(n)} = \sup_{t\in I_n} Z_t$. Then $\left(Z_T^{(n)}\right)_{n=1}^{\infty}$ is a non-decreasing sequence, and $Z_T^{(n)} \to Z_T^*$ as $n \to \infty$, almost surely. For each $Z_T^{(n)}$, we can apply Lemma 4.4.1, and it is immediate that for any $b > 0$,

$$\mathbb{P}(Z_T^{(n)} > b) \leq \frac{BT}{b}, \tag{4.9}$$

since each I_n includes both 0 and T. Since $Z_T^{(n)}$ increases monotonically to Z_T^*, almost surely, we have that $\mathbb{P}(Z_T^{(n)} > b) \leq \mathbb{P}(Z_T^{(n+1)} > b)$ for all n, and $\mathbb{P}(Z_T^{(n)} > b) \to \mathbb{P}(Z_T^* > b)$ as $n \to \infty$. The right-hand side of (4.9) is fixed, so

$$\mathbb{P}(Z_T^* > b) \leq \frac{BT}{b}.$$

We now take an increasing sequence b_n with $\lim_{n\to\infty} b_n = a$, and obtain

$$\mathbb{P}(Z_T^* \geq a) \leq \frac{BT}{a}.$$

□

4.4.2 A Maximal Inequality for Bandwidth-Sharing Networks

We employ the Lyapunov function (4.7) to study α-fair policies. This is the Lyapunov function that was used in [6], [15] and [36] to establish positive recurrence of the process $\mathbf{M}(\cdot)$ under an α-fair policy. Below we fine-tune the proof in [15] to obtain a more precise bound on the Lyapunov drift. We note that a "fluid-model" version of the following lemma appeared in the proof of Theorem 1 in [6]. For notational convenience, we drop the subscript α from F_α and write F instead.

Lemma 4.4.3 *Consider a bandwidth-sharing network with $\mathbf{R}\boldsymbol{\lambda} < \mathbf{C}$ operating under an α-fair policy with $\alpha > 0$. Let ε be the gap. Then for any non-zero flow vector $\mathbf{m}$,*

$$\langle \nabla F(\mathbf{m}), \boldsymbol{\nu} - \boldsymbol{\mu}\boldsymbol{\phi}(\mathbf{m}) \rangle \leq -\varepsilon \langle \nabla F(\mathbf{m}), \boldsymbol{\nu} \rangle ,$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product, $\nabla F(\mathbf{m})$ denotes the gradient of F, and $\boldsymbol{\mu}\boldsymbol{\phi}(\mathbf{m})$ is the vector $(\mu_i \phi_i(\mathbf{m}))_{i\in\mathcal{I}}$.

Proof. We have

$$\begin{aligned}
\langle \nabla F(\mathbf{m}), \boldsymbol{\nu} - \boldsymbol{\mu}\boldsymbol{\phi}(\mathbf{m}) \rangle &= \sum_{i\in\mathcal{I}} \frac{1}{\mu_i} \kappa_i \left(\frac{m_i}{\lambda_i}\right)^\alpha (\nu_i - \mu_i \phi_i(\mathbf{m})) \\
&= \sum_{i\in\mathcal{I}} \kappa_i \left(\frac{m_i}{\lambda_i}\right)^\alpha (\lambda_i - \phi_i(\mathbf{m})) \\
&= \left\langle \nabla G_{\mathbf{m}}(\boldsymbol{\lambda}_+), \boldsymbol{\lambda}_+ - \boldsymbol{\phi}_+(\mathbf{m}) \right\rangle ,
\end{aligned}$$

where $\boldsymbol{\lambda}_+ = (\lambda_i)_{i\in\mathcal{I}_+(\mathbf{m})}$. Similarly we can get $\langle \nabla F(\mathbf{m}), \boldsymbol{\nu} \rangle = \langle \nabla G_{\mathbf{m}}(\boldsymbol{\lambda}_+), \boldsymbol{\lambda}_+ \rangle$.

Now consider the function $g : [0, 1] \to \mathbb{R}$ defined by

$$g(\theta) = G_{\mathbf{m}}\big(\theta(1+\varepsilon)\boldsymbol{\lambda}_+ + (1-\theta)\boldsymbol{\phi}_+(\mathbf{m})\big).$$

Since $(1+\varepsilon)\boldsymbol{\lambda}_+$ satisfies the constraints in (4.2), and $\boldsymbol{\phi}_+(\mathbf{m})$ maximizes the strictly *concave* function $G_{\mathbf{m}}$ subject to the constraints in (4.2), we have

$$G_{\mathbf{m}}((1+\varepsilon)\boldsymbol{\lambda}_+) \leq G_{\mathbf{m}}(\boldsymbol{\phi}_+(\mathbf{m})), \quad \text{i.e., } g(1) \leq g(0).$$

Furthermore, since $G_{\mathbf{m}}$ is a concave function, g is also concave in θ. Thus,

$$g(0) \leq g(1) + (0-1)g'(1) \leq g(0) + (0-1)g'(1).$$

Hence, $g'(1) \leq 0$, i.e.,

$$\frac{dg}{d\theta}\Big|_{\theta=1} = \left\langle \nabla G_{\mathbf{m}}((1+\varepsilon)\boldsymbol{\lambda}_+), (1+\varepsilon)\boldsymbol{\lambda}_+ - \boldsymbol{\phi}_+(\mathbf{m}) \right\rangle \leq 0. \tag{4.10}$$

But it is easy to check that $\nabla G_{\mathbf{m}}((1+\varepsilon)\boldsymbol{\lambda}_+) = (1+\varepsilon)^{-\alpha}\nabla G_{\mathbf{m}}(\boldsymbol{\lambda}_+)$, so dividing (4.10) by $(1+\varepsilon)^{-\alpha}$, we have

$$\left\langle \nabla G_{\mathbf{m}}(\boldsymbol{\lambda}_+), \boldsymbol{\lambda}_+ - \boldsymbol{\phi}_+(\mathbf{m}) \right\rangle \leq -\varepsilon \left\langle \nabla G_{\mathbf{m}}(\boldsymbol{\lambda}_+), \boldsymbol{\lambda}_+ \right\rangle.$$

This is the same as

$$\left\langle \nabla F(\mathbf{m}), \boldsymbol{\nu} - \boldsymbol{\mu}\boldsymbol{\phi}(\mathbf{m}) \right\rangle \leq -\varepsilon \left\langle \nabla F(\mathbf{m}), \boldsymbol{\nu} \right\rangle. \qquad \square$$

Our next lemma provides a uniform upper bound on the expected change of $F(\tilde{\mathbf{M}}(\cdot))$ in one time step, where $\tilde{\mathbf{M}}(\cdot)$ is the uniformized chain associated with the Markov process $\mathbf{M}(\cdot)$ (cf. Definition 4.2.3).

Lemma 4.4.4 *Let $\alpha \geq 1$. As above, consider a bandwidth-sharing network with $\mathbf{R}\boldsymbol{\lambda} < \mathbf{C}$ operating under an α-fair policy. Let ε be the gap. Let $\left(\tilde{\mathbf{M}}(\tau)\right)_{\tau\in\mathbb{Z}_+}$ be the uniformized chain associated with the Markov process $\mathbf{M}(\cdot)$. Then, there exists a positive load-dependent constant $\bar{K}$, such that for all $\tau \in \mathbb{Z}_+$,*

$$\mathbb{E}\left[F(\tilde{\mathbf{M}}(\tau+1)) - F(\mathbf{m}) \mid \tilde{\mathbf{M}}(\tau) = \mathbf{m}\right] \leq \bar{K}\varepsilon^{1-\alpha}.$$

Proof. By the mean value theorem (cf. Proposition 3.3.1), for $\mathbf{m}, \in \mathbb{Z}_+^N$, we have

$$F(\mathbf{m}+\mathbf{n}) - F(\mathbf{m}) = \langle \nabla F(\mathbf{m}), \mathbf{n} \rangle + \frac{1}{2}\mathbf{n}^T \nabla^2 F(\mathbf{m}+\theta\mathbf{n})\mathbf{n}, \tag{4.11}$$

for some $\theta \in [0,1]$. We note that, for $\mathbf{n} = \pm\mathbf{e}_i$, we have

$$\begin{aligned} \frac{1}{2}\mathbf{n}^T \nabla^2 F(\mathbf{m}+\theta\mathbf{n})\mathbf{n} &\leq \frac{\kappa_i \alpha}{2\mu_i \lambda_i^\alpha}(m_i \pm \theta)^{\alpha-1} \\ &\leq \frac{\kappa_i \alpha}{2\mu_i \lambda_i^\alpha}(m_i+1)^{\alpha-1}, \end{aligned} \tag{4.12}$$

since $\alpha \geq 1$, and $\theta \in [0,1]$.

As in [15], we define

$$\mathbf{G}F(\mathbf{m}) \triangleq \sum_{\mathbf{n}} q(\mathbf{m}, \mathbf{m}+\mathbf{n})[F(\mathbf{m}+\mathbf{n}) - F(\mathbf{m})],$$

so that $\mathbf{G}$ is the generator of the Markov process $\mathbf{M}(\cdot)$. We now proceed to derive an upper bound for $\mathbf{G}F(\mathbf{m})$. Using Equation (4.11), we can rewrite $\mathbf{G}F(\mathbf{m})$ as

$$\begin{aligned} \mathbf{G}F(\mathbf{m}) &= \sum_{\mathbf{n}} q(\mathbf{m}, \mathbf{m}+\mathbf{n}) \left[\langle \nabla F(\mathbf{m}), \mathbf{n} \rangle + \frac{1}{2}\mathbf{n}^T \nabla^2 F(\mathbf{m}+\theta_{\mathbf{n}}\mathbf{n})\mathbf{n} \right] \\ &= \sum_{\mathbf{n}} q(\mathbf{m}, \mathbf{m}+\mathbf{n}) \langle \nabla F(\mathbf{m}), \mathbf{n} \rangle \\ &\quad + \frac{1}{2}\sum_{\mathbf{n}} q(\mathbf{m}, \mathbf{m}+\mathbf{n})\mathbf{n}^T \nabla^2 F(\mathbf{m}+\theta_{\mathbf{n}}\mathbf{n})\mathbf{n}, \end{aligned}$$

for some scalars $\theta_{\mathbf{n}} \in [0,1]$, one such scalar for each $\mathbf{n}$. From the definition of $\mathbf{q}$, we have

$$\begin{aligned} \sum_{\mathbf{n}} q(\mathbf{m}, \mathbf{m}+\mathbf{n}) \langle \nabla F(\mathbf{m}), \mathbf{n} \rangle &= \left\langle \nabla F(\mathbf{m}), \sum_{\mathbf{n}} q(\mathbf{m}, \mathbf{m}+\mathbf{n})\mathbf{n} \right\rangle \\ &= \langle \nabla F(\mathbf{m}), \boldsymbol{\nu} - \boldsymbol{\mu}\boldsymbol{\phi}(\mathbf{m}) \rangle . \end{aligned}$$

From (4.12), for $\mathbf{n} = \pm\mathbf{e}_i$, we also have

$$\frac{1}{2}\mathbf{n}^T \nabla^2 F(\mathbf{m}+\theta_{\mathbf{n}}\mathbf{n})\mathbf{n} \leq \kappa_i \alpha (m_i+1)^{\alpha-1}/2\mu_i \lambda_i^\alpha.$$

Thus,

$$\begin{aligned}
\mathbf{G}F(\mathbf{m}) &\leq \langle \nabla F(\mathbf{m}), \boldsymbol{\nu} - \boldsymbol{\mu}\boldsymbol{\phi}(\mathbf{m}) \rangle + \sum_{i \in \mathcal{I}} \frac{\kappa_i \alpha}{2\mu_i \lambda_i^{\alpha}} (m_i + 1)^{\alpha - 1} (\nu_i + \mu_i \phi_i(\mathbf{m})) \\
&\leq -\varepsilon \sum_{i \in \mathcal{I}} \kappa_i \left(\frac{m_i}{\lambda_i} \right)^{\alpha} \lambda_i + \sum_{i \in \mathcal{I}} \frac{\kappa_i \alpha}{2\lambda_i^{\alpha}} (m_i + 1)^{\alpha - 1} (\lambda_i + \phi_i(\mathbf{m})) \\
&\leq -\gamma \varepsilon \sum_{i \in I} m_i^{\alpha} + \Gamma \sum_{i \in I} (m_i + 1)^{\alpha - 1},
\end{aligned}$$

where the second inequality follows from Lemma 4.4.3, and the third by defining

$$\gamma \triangleq \min_{i \in \mathcal{I}} \kappa_i \lambda_i^{1-\alpha}, \qquad \Gamma \triangleq \max_{i \in \mathcal{I}} \frac{\kappa_i \alpha}{2\lambda_i^{\alpha}} \left(\lambda_i + \max_{j \in \mathcal{J}} C_j \right),$$

and noting the fact that since $\phi_i(\mathbf{m}) \leq \max_{j \in \mathcal{J}} C_j$ for all i, we have $\Gamma \geq \max_{i \in \mathcal{I}} \frac{\kappa_i \alpha}{2\lambda_i^{\alpha}} \big(\lambda_i + \phi_i(\mathbf{m}) \big)$. It is then a simple calculation to see that for every $\mathbf{m} \geq \mathbf{0}$, we have

$$\mathbf{G}F(\mathbf{m}) \leq -\gamma \varepsilon \sum_{i \in I} m_i^{\alpha} + \Gamma \sum_{i \in I} (m_i + 1)^{\alpha - 1} \leq \tilde{K} \varepsilon^{1-\alpha},$$

for some positive load-dependent constant $\tilde{K}$. Now given $\tilde{\mathbf{M}}(\tau) = \mathbf{m}$,

$$\mathbb{E}\left[F(\tilde{\mathbf{M}}(\tau + 1)) - F(\mathbf{m}) \mid \tilde{\mathbf{M}}(\tau) = \mathbf{m} \right] = \frac{\mathbf{G}F(\mathbf{m})}{\Xi} \leq \frac{\tilde{K} \varepsilon^{1-\alpha}}{\Xi}.$$

By setting $\bar{K} = \tilde{K}/\Xi$, we have proved the lemma. □

Corollary 4.4.5 *Let $\alpha \geq 1$. As before, suppose that $\mathbf{R}\boldsymbol{\lambda} < \mathbf{C}$, and let ε be the associated gap. Then, under the α-fair policy, the process $\mathbf{M}(\cdot)$ satisfies*

$$\mathbb{E}\left[F(\mathbf{M}(s + t)) - F(\mathbf{M}(s)) \mid \mathbf{M}(s) \right] \leq \tilde{K} t \varepsilon^{1-\alpha}, \quad \textit{for all } t \geq 0.$$

for some positive load-dependent constant $\tilde{K}$.

Proof. The idea of the proof is to show that the expected number of state transitions of $\mathbf{M}(\cdot)$ in the time interval $[s, s + t]$ is of order $O(t)$.

Consider the uniformized Markov chain $\tilde{\mathbf{M}}(\cdot)$ associated with the process $\mathbf{M}(\cdot)$.

Denote the number of state transitions in the uniformized version of the process $\mathbf{M}(\cdot)$ in the time interval $[s, s+t]$ by τ. By the Markov property, time-homogeneity, and the definition of $\tilde{\mathbf{M}}(\cdot)$, we have

$$\begin{aligned} & \mathbb{E}\left[F(\mathbf{M}(s+t)) - F(\mathbf{M}(s)) \mid \mathbf{M}(s) = \mathbf{m}\right] \\ = \; & \mathbb{E}\left[F(\tilde{\mathbf{M}}(\tau)) - F(\tilde{\mathbf{M}}(0)) \mid \tilde{\mathbf{M}}(0) = \mathbf{m}\right]. \end{aligned}$$

Now, by the definition of the uniformized chain, τ and $\tilde{\mathbf{M}}(\cdot)$ are independent. Thus,

$$\begin{aligned} & \mathbb{E}\left[F(\tilde{\mathbf{M}}(\tau)) - F(\tilde{\mathbf{M}}(0)) \mid \tilde{\mathbf{M}}(0)\right] \\ = \; & \mathbb{E}\left[\sum_{k=0}^{\tau-1}\left(F(\tilde{\mathbf{M}}(k+1)) - F(\tilde{\mathbf{M}}(k))\right) \;\middle|\; \tilde{\mathbf{M}}(0)\right] \\ = \; & \mathbb{E}\left[\mathbb{E}\left[\sum_{k=0}^{\tau-1}\left(F(\tilde{\mathbf{M}}(k+1)) - F(\tilde{\mathbf{M}}(k))\right) \;\middle|\; \tilde{\mathbf{M}}(0), \tau\right] \;\middle|\; \tilde{\mathbf{M}}(0)\right] \\ = \; & \mathbb{E}\left[\sum_{k=0}^{\tau-1}\mathbb{E}\left[F(\tilde{\mathbf{M}}(k+1)) - F(\tilde{\mathbf{M}}(k)) \;\middle|\; \tilde{\mathbf{M}}(0), \tau\right] \;\middle|\; \tilde{\mathbf{M}}(0)\right] \\ = \; & \mathbb{E}\left[\sum_{k=0}^{\tau-1}\mathbb{E}\left[F(\tilde{\mathbf{M}}(k+1)) - F(\tilde{\mathbf{M}}(k)) \mid \tilde{\mathbf{M}}(0)\right] \;\middle|\; \tilde{\mathbf{M}}(0)\right] \\ \leq \; & \mathbb{E}\left[\sum_{k=0}^{\tau-1}\bar{K}\varepsilon^{1-\alpha}\right] = \bar{K}\varepsilon^{1-\alpha}\mathbb{E}[\tau], \end{aligned}$$

for some load-dependent constant $\bar{K}$. The fourth equality follows from the independence of τ and $\tilde{\mathbf{M}}(\cdot)$, and the inequality follows from Lemma 4.4.4. Since the counting process of the number of state transitions in the uniformized version of the process $\mathbf{M}(\cdot)$ is a time-homogeneous Poisson process of rate Ξ, we have $\mathbb{E}[\tau] = \Xi t$. This shows that

$$\mathbb{E}\left[F(\mathbf{M}(s+t)) - F(\mathbf{M}(s)) \mid \mathbf{M}(s)\right] \leq \bar{K}\Xi t\varepsilon^{1-\alpha}.$$

The proof is concluded by setting $\tilde{K} = \bar{K}\Xi$. □

Proof of Theorem 4.3.1. Let $b > 0$. Then

$$\begin{aligned}
\mathbb{P}\left(N^*(T) \geq b\right) &= \mathbb{P}\left(\frac{1}{\alpha+1}\left(N^*(T)\right)^{\alpha+1} \geq \frac{1}{\alpha+1}b^{\alpha+1}\right) \\
&\leq \mathbb{P}\left(\sup_{t\in[0,T]} F(\mathbf{M}(t)) \geq \left(\min_{i\in\mathcal{I}} \frac{1}{\alpha+1}\kappa_i \mu_i^{\alpha-1}\nu_i^{-\alpha}\right)b^{\alpha+1}\right) \\
&\leq \frac{(\alpha+1)K'T}{\left(\min_{i\in\mathcal{I}} \kappa_i \mu_i^{\alpha-1}\nu_i^{-\alpha}\right)\varepsilon^{\alpha-1}b^{\alpha+1}} = \frac{KT}{\varepsilon^{\alpha-1}b^{\alpha+1}},
\end{aligned}$$

where the second inequality follows from Corollary 4.4.2 and Corollary 4.4.5, K' is as in Corollary 4.4.5, and $K = \frac{(\alpha+1)K'}{\min_{i\in\mathcal{I}} \kappa_i \mu_i^{\alpha-1}\nu_i^{-\alpha}}$.

4.4.3 Full State Space Collapse for $\alpha \geq 1$

Throughout this section, we assume that we have fixed $\alpha \geq 1$, and correspondingly, the Lyapunov function (4.7). To state the full state space collapse result for $\alpha \geq 1$, we need some preliminary definitions and the statement of the multiplicative state space collapse result.

Consider a sequence of bandwidth-sharing networks indexed by r, where r is to be thought of as increasing to infinity along a sequence. Suppose that the incidence matrix $\mathbf{R}$, the capacity vector $\mathbf{C}$ and the weights $\{\kappa_i : i \in \mathcal{I}\}$ do not vary with r. Write $\mathbf{M}^r(t)$ for the flow-vector Markov process associated with the rth network. Similarly, we write $\boldsymbol{\nu}^r$, $\boldsymbol{\mu}^r$, $\boldsymbol{\lambda}^r$, etc. We assume the following *heavy-traffic* condition (cf. [32]):

Assumption 4.4.6 *We assume that* $\mathbf{R}\boldsymbol{\lambda}^r < \mathbf{C}$ *for all* r. *We also assume that there exist* $\boldsymbol{\nu}, \boldsymbol{\mu} \in \mathbb{R}_+^N$ *and* $\boldsymbol{\theta} > \mathbf{0}$, *such that* $\nu_i > 0$ *and* $\mu_i > 0$ *for all* $i \in \mathcal{I}$, $\boldsymbol{\nu}^r \to \boldsymbol{\nu}$ *and* $\boldsymbol{\mu}^r \to \boldsymbol{\mu}$ *as* $r \to \infty$, *and* $r(\mathbf{C} - \mathbf{R}\boldsymbol{\lambda}^r) \to \boldsymbol{\theta}$ *as* $r \to \infty$.

Note that our assumption differs from that in [32], which allows convergence to the critical load from both overload and underload, whereas here we only allow convergence to the critical load from underload.

To state the multiplicative state space collapse result, we also need to define a *workload process* $\mathbf{W}(t)$ and a *lifting map* Δ.

Definition 4.4.7 *We first define the* workload $\mathbf{w} : \mathbb{R}_+^N \to \mathbb{R}_+^J$ *associated with a flow-vector* $\mathbf{m}$ *by* $\mathbf{w} = \mathbf{w}(\mathbf{m}) = \mathbf{R}\mathbf{E}^{-1}\mathbf{m}$, *where* $\mathbf{E} = \mathrm{diag}(\boldsymbol{\mu})$ *is the* $N \times N$ *diagonal matrix with* $\boldsymbol{\mu}$ *on its diagonal. The* workload process $\mathbf{W}(t)$ *is defined to be* $\mathbf{W}(t) \triangleq \mathbf{R}\mathbf{E}^{-1}\mathbf{M}(t)$, *for all* $t \geq 0$. *We also define the* lifting map Δ. *For each* $\mathbf{w} \in \mathbb{R}_+^J$, *define* $\Delta(\mathbf{w})$ *to be the unique value of* $\mathbf{m} \in \mathbb{R}_+^N$ *that solves the following optimization problem:*

$$\begin{array}{ll} \textit{minimize} & F(\mathbf{m}) \\ \textit{subject to} & \sum_{i\in\mathcal{I}} R_{ji}\frac{m_i}{\mu_i} \geq w_j, \quad j \in \mathcal{J}, \\ & m_i \geq 0, \quad i \in \mathcal{I}. \end{array}$$

For simplicity, suppose that all networks start with zero flows. We consider the following diffusion scaling:

$$\hat{\mathbf{M}}^r(t) = \frac{\mathbf{M}^r(r^2t)}{r}, \text{ and } \hat{\mathbf{W}}^r(t) = \frac{\mathbf{W}^r(r^2t)}{r}, \tag{4.13}$$

where $\mathbf{W}^r(t) = \mathbf{R}(\mathbf{E}^r)^{-1}\mathbf{M}^r(t)$, and $\mathbf{E}^r = \mathrm{diag}(\boldsymbol{\mu}^r)$.

The following multiplicative state space collapse result is known to hold.

Theorem 4.4.8 (Multiplicative State Space Collapse [32, Theorem 5.1]) *Fix* $T > 0$ *and assume that* $\alpha \geq 1$. *Write* $\|\mathbf{x}(\cdot)\| = \sup_{t\in[0,T],i\in\mathcal{I}} |x_i(t)|$. *Then, under Assumption 4.4.6, and for any* $\delta > 0$,

$$\lim_{r\to\infty} \mathbb{P}\left(\frac{\|\hat{\mathbf{M}}^r(\cdot) - \Delta(\hat{\mathbf{W}}^r(\cdot))\|}{\|\hat{\mathbf{M}}^r(\cdot)\|} > \delta\right) = 0.$$

We can now state and prove a full state space collapse result:

Theorem 4.4.9 (Full State Space Collapse) *Under the same assumptions as in Theorem 4.4.8, and for any* $\delta > 0$,

$$\lim_{r\to\infty} \mathbb{P}\Big(\|\hat{\mathbf{M}}^r(\cdot) - \Delta(\hat{\mathbf{W}}^r(\cdot))\| > \delta\Big) = 0.$$

Proof. Let $\varepsilon_r = \varepsilon(\boldsymbol{\lambda}^r)$ be the gap in the rth system. Then, under Assumption 4.4.6, $\varepsilon_r \geq D/r$ for some network-dependent constant $D > 0$, and for r sufficiently large.

By Theorem 4.3.1, for any $b > 0$, and for sufficiently large r,

$$\begin{aligned}\mathbb{P}\left(N^{r,*}(r^2T) \geq b\right) &\leq \frac{K_r r^2 T}{\varepsilon_r^{\alpha-1} b^{\alpha+1}} \\ &\leq \frac{K_r r^{1+\alpha} T}{D^{\alpha-1} b^{\alpha+1}}.\end{aligned}$$

Here, K_r is a load-dependent constant associated with the rth system, as specified in the proof of Theorem 4.3.1. From the proof of Theorem 4.3.1, note also that $K_r = f(\boldsymbol{\mu}^r, \boldsymbol{\nu}^r)$, for a function f that is continuous on the open positive orthant $\mathrm{R}_p^N \times \mathrm{R}_p^N$. Since $\boldsymbol{\mu}^r \to \boldsymbol{\mu} > \mathbf{0}$, and $\boldsymbol{\nu}^r \to \boldsymbol{\nu} > \mathbf{0}$, $K_r \to K \triangleq f(\boldsymbol{\mu}, \boldsymbol{\nu}) \in \mathrm{R}$. In particular, the K_r are bounded, and for all sufficiently large r,

$$\mathbb{P}\left(N^{r,*}(r^2T) \geq b\right) \leq \frac{(K+1) r^{1+\alpha} T}{D^{\alpha-1} b^{\alpha+1}}.$$

Then, with $a = b/r$ and under the scaling in (4.13),

$$\mathbb{P}\big(\|\hat{\mathbf{N}}^r(\cdot)\| \geq a\big) \leq \frac{K+1}{D^{\alpha-1}} \cdot \frac{T}{a^{\alpha+1}}, \tag{4.14}$$

for any $a > 0$.

For notational convenience, we write

$$B(r) = \|\hat{\mathbf{N}}^r(\cdot) - \Delta(\hat{\mathbf{W}}^r(\cdot))\|.$$

Then, for any $a > 1$, and for sufficiently large r,

$$\begin{aligned}\mathbb{P}\big(B(r) > \delta\big) &\leq \mathbb{P}\left(\frac{B(r)}{\|\hat{\mathbf{N}}^r(\cdot)\|} > \frac{\delta}{a} \text{ or } \|\hat{\mathbf{N}}^r(\cdot)\| \geq a\right) \\ &\leq \mathbb{P}\left(\frac{B(r)}{\|\hat{\mathbf{N}}^r(\cdot)\|} > \frac{\delta}{a}\right) + \mathbb{P}\big(\|\hat{\mathbf{N}}^r(\cdot)\| \geq a\big).\end{aligned}$$

Note that by Theorem 4.4.8, the first term on the right-hand side goes to 0 as $r \to \infty$, for any $a > 0$. The second term on the right-hand side can be made smaller than any, arbitrarily small, constant (uniformly, for all r), by taking a sufficiently large (cf.

Equation (4.14)). Thus, $\mathbb{P}(B(r) \geq \delta) \to 0$ as $r \to \infty$. This concludes the proof. $\square$

4.5 Steady-State Analysis ($\alpha > 0$)

4.5.1 α-Fair Policies: A Useful Drift Inequality

We now shift our focus to the steady-state regime. As in Section 3.4.1, the key to many of our results is a *drift inequality* that holds for every $\alpha > 0$ and every $\boldsymbol{\lambda} > \mathbf{0}$ with $\mathbf{R}\boldsymbol{\lambda} < \mathbf{C}$. In this section, we shall state and prove this inequality. It will be used in Section 4.5.2 to prove Theorem 4.3.2.

We define the Lyapunov function that we will employ. It will be very similar to the Lyapunov function used in Section 3.4.1, Chapter 3. For $\alpha \geq 1$, it will be simply the weighted $(\alpha+1)$-norm $L_\alpha(\mathbf{m}) = \sqrt[\alpha+1]{(\alpha+1)F_\alpha(\mathbf{m})}$ of a vector $\mathbf{m}$, where F_α was defined in (4.7). However, when $\alpha \in (0,1)$, this function has unbounded second derivatives as we approach the boundary of $\mathbb{R}^N_+$. For this reason, our Lyapunov function will be a suitably smoothed version of $\sqrt[\alpha+1]{(\alpha+1)F_\alpha(\cdot)}$. As in Section 3.4.1, Chapter 3, we will make use of the following functions h_α and H_α to define our Lyapunov functions, and Lemma 4.5.2 is an exact copy of Lemma 3.4.2, included for easy reference.

Definition 4.5.1 *Define $h_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ to be $h_\alpha(r) = r^\alpha$, when $\alpha \geq 1$, and*

$$h_\alpha(r) = \begin{cases} r^\alpha, & \text{if } r \geq 1, \\ (\alpha-1)r^3 + (1-\alpha)r^2 + r, & \text{if } r < 1, \end{cases}$$

when $\alpha \in (0,1)$. Let $H_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ be the antiderivative of h_α, so that $H_\alpha(r) = \int_0^r h_\alpha(s)\,ds$. The Lyapunov function $L_\alpha : \mathbb{R}^N_+ \to \mathbb{R}_+$ is defined to be

$$L_\alpha(\mathbf{n}) = \left[(\alpha+1)\sum_{i\in\mathcal{I}} \kappa_i \mu_i^{\alpha-1} \nu_i^{-\alpha} H_\alpha(m_i)\right]^{\frac{1}{\alpha+1}}.$$

For notational convenience, define

$$w_i = \kappa_i \mu_i^{\alpha-1} \nu_i^{-\alpha} \text{ for each } i \in \mathcal{I}, \tag{4.15}$$

so that more compactly, we have

$$F_\alpha(\mathbf{m}) = \frac{1}{\alpha+1} \sum_{i \in \mathcal{I}} w_i m_i^{\alpha+1}, \text{ and } L_\alpha(\mathbf{m}) = \left[(\alpha+1) \sum_{i \in \mathcal{I}} w_i H_\alpha(m_i) \right]^{1/(\alpha+1)}.$$

Lemma 4.5.2 *Let $\alpha \in (0,1)$. The function h_α has the following properties:*

(i) it is continuously differentiable with $h_\alpha(0) = 0$, $h_\alpha(1) = 1$, $h'_\alpha(0) = 1$, and $h'_\alpha(1) = \alpha$;

(ii) it is increasing and, in particular, $h_\alpha(r) \geq 0$ for all $r \geq 0$;

(iii) we have $r^\alpha - 1 \leq h_\alpha(r) \leq r^\alpha + 1$, for all $r \in [0,1]$;

(iv) $h'_\alpha(r) \leq 2$, for all $r \geq 0$.

Furthermore, from (iii), we also have the following property of H_α:

(iii') $r^{\alpha+1} - 2 \leq (\alpha+1) H_\alpha(r) \leq r^{\alpha+1} + 2$ for all $r \geq 0$.

We are now ready to state the drift inequality. Here we consider the uniformized chain $\left(\tilde{\mathbf{M}}(\tau)\right)_{\tau \in \mathbb{Z}_+}$ associated with $\mathbf{M}(\cdot)$, and the corresponding drift.

Theorem 4.5.3 *Consider a bandwidth-sharing network operating under an α-fair policy with $\alpha > 0$, and assume that $\mathbf{R}\boldsymbol{\lambda} < \mathbf{C}$. Let ε be the gap. Then, there exists a positive constant B and a positive load-dependent constant K, such that if $L_\alpha(\tilde{\mathbf{M}}(\tau)) > B$, then*

$$\mathbb{E}[L_\alpha(\tilde{\mathbf{M}}(\tau+1)) - L_\alpha(\tilde{\mathbf{M}}(\tau)) \mid \tilde{\mathbf{M}}(\tau)] \leq -\varepsilon K. \tag{4.16}$$

Furthermore, B takes the form K'/ε when $\alpha \geq 1$, and $K'/\min\{\varepsilon^{1/\alpha}, \varepsilon\}$ when $\alpha \in (0,1)$, with K' being a positive load-dependent constant.

As there is a marked difference between the form of L_α for the two cases $\alpha \geq 1$ and $\alpha \in (0,1)$, the proof of the drift inequality is split into two parts. We first prove the drift inequality when $\alpha \geq 1$, in which case L_α takes a nicer form, and we can apply results on F_α from previous sections. The proof for the case $\alpha \in (0,1)$ is similar but more tedious. We note that such a qualitative difference between the two cases, $\alpha < 1$ and $\alpha \geq 1$, has also been observed in other works, such as, for example, [56].

We wish to draw attention here to the main difference from related drift inequalities in the literature. The usual proof of stability involves the Lyapunov function (4.7); for instance, for the α-fair policy with $\alpha = 1$ (the proportionally fair policy), it involves a weighted quadratic Lyapunov function. In contrast, we use L_α, a weighted norm function (or its smoothed version), which scales linearly along radial directions. In this sense, our approach is similar in spirit to [4], which employed piecewise linear Lyapunov functions to derive drift inequalities and then moment and tail bounds. The use of normed Lyapunov functions to establish stability and performance bounds has also been considered in other works; see, for example, [62] and [17].

Proof of Theorem 4.5.3: $\alpha \geq 1$. We first consider the case $\alpha \geq 1$. We wish to decompose the drift term in (4.16) into the sum of a first-order term and a second-order term, and we accomplish this by using the second-order mean value theorem (cf. Proposition 3.3.1). Throughout this proof, we drop the subscript α from L_α and F_α, and write L and F, respectively.

Consider the function $L(\mathbf{n}) = \left(\sum_{i\in\mathcal{I}} w_i m_i^{\alpha+1}\right)^{\frac{1}{\alpha+1}} = [(\alpha+1)F(\mathbf{m})]^{\frac{1}{\alpha+1}}$. The first derivative of L with respect to $\mathbf{m}$ is $\nabla L(\mathbf{n}) = \nabla F(\mathbf{m})/L^\alpha(\mathbf{m})$ by the chain rule and the definition of L. The second derivative is

$$\begin{aligned}\nabla^2 L(\mathbf{m}) &= \frac{\nabla^2 F(\mathbf{m})}{L^\alpha(\mathbf{m})} - \frac{\nabla F(\mathbf{m})\nabla L^\alpha(\mathbf{m})^T}{L^{2\alpha}(\mathbf{m})} \\ &= \frac{\nabla^2 F(\mathbf{m})}{L^\alpha(\mathbf{m})} - \alpha\frac{\nabla F(\mathbf{m})\nabla F(\mathbf{m})^T}{L^{2\alpha+1}(\mathbf{m})},\end{aligned}$$

by the quotient rule and the chain rule.

Write $\mathbf{m}$ for $\tilde{\mathbf{M}}(\tau)$ and $\mathbf{m}+\mathbf{n}$ for $\tilde{\mathbf{M}}(\tau+1)$, so that $\mathbf{n} = \tilde{\mathbf{M}}(\tau+1) - \tilde{\mathbf{M}}(\tau)$. By

Proposition 3.3.1, for some $\theta \in [0,1]$, we have

$$\begin{aligned} L(\mathbf{m}+\mathbf{n}) - L(\mathbf{m}) &= \mathbf{n}^T \nabla L(\mathbf{m}) + \frac{1}{2}\mathbf{n}^T \nabla^2 L(\mathbf{m}+\theta\mathbf{n})\mathbf{n} && (4.17) \\ &= \frac{\mathbf{n}^T \nabla F(\mathbf{m})}{L^\alpha(\mathbf{m})} + \frac{1}{2}\frac{\mathbf{n}^T \nabla^2 F(\mathbf{m}+\theta\mathbf{n})\mathbf{n}}{L^\alpha(\mathbf{m}+\theta\mathbf{n})} && (4.18) \\ &\quad -\frac{\alpha}{2}\frac{\mathbf{n}^T \nabla F(\mathbf{m}+\theta\mathbf{n})\nabla F(\mathbf{m}+\theta\mathbf{n})^T\mathbf{n}}{L^{2\alpha+1}(\mathbf{m}+\theta\mathbf{n})} && (4.19) \\ &\leq \frac{\mathbf{n}^T \nabla F(\mathbf{m})}{L^\alpha(\mathbf{m})} + \frac{1}{2}\mathbf{n}^T \frac{\nabla^2 F(\mathbf{m}+\theta\mathbf{n})}{L^\alpha(\mathbf{m}+\theta\mathbf{n})}\mathbf{n}, && (4.20) \end{aligned}$$

since the term $\mathbf{n}^T \nabla F(\mathbf{m}+\theta\mathbf{n})\nabla F(\mathbf{m}+\theta\mathbf{n})^T\mathbf{n}$ is nonnegative. We now consider the two terms in (4.20) separately. Recall from the proof of Lemma 4.4.4 that

$$\mathbb{E}\left[\mathbf{n}^T \nabla F(\mathbf{m}) \mid \mathbf{m}\right] = \frac{\langle \nabla F(\mathbf{m}), \boldsymbol{\nu} - \boldsymbol{\mu}\boldsymbol{\phi}(\mathbf{m})\rangle}{\Xi} \leq -\varepsilon \frac{\langle \nabla F(\mathbf{m}), \boldsymbol{\nu}\rangle}{\Xi}.$$

But $\langle \nabla F(\mathbf{m}), \boldsymbol{\nu}\rangle = \sum_{i\in\mathcal{I}} w_i \nu_i m_i^\alpha$, so

$$\mathbb{E}\left[\mathbf{n}^T \nabla F(\mathbf{m}) \mid \mathbf{m}\right] \leq -\varepsilon \frac{\sum_{i\in\mathcal{I}} w_i \nu_i m_i^\alpha}{\Xi}, \tag{4.21}$$

and so

$$\begin{aligned} \mathbb{E}\left[\frac{\mathbf{n}^T \nabla F(\mathbf{m})}{L^\alpha(\mathbf{m})} \,\middle|\, \mathbf{m}\right] &\leq -\varepsilon \frac{\sum_{i\in\mathcal{I}} w_i \nu_i m_i^\alpha}{\Xi \left(\sum_{i\in\mathcal{I}} w_i m_i^{\alpha+1}\right)^{\frac{\alpha}{\alpha+1}}} \\ &= -\varepsilon \frac{\sum_{i\in\mathcal{I}} w_i \nu_i m_i^\alpha}{\Xi \left(\sum_{i\in\mathcal{I}} \left(w_i^{\frac{1}{\alpha+1}} m_i\right)^{\alpha+1}\right)^{\frac{\alpha}{\alpha+1}}} \\ &\leq -\varepsilon \frac{\sum_{i\in\mathcal{I}} w_i \nu_i m_i^\alpha}{\Xi \cdot \sum_{i\in I} w_i^{\frac{\alpha}{\alpha+1}} m_i^\alpha} \\ &\leq -\varepsilon \frac{\max_{i\in\mathcal{I}} w_i^{\frac{1}{\alpha+1}} \nu_i}{\Xi} \\ &= -\varepsilon \frac{\max_{i\in\mathcal{I}} \kappa^{\frac{1}{\alpha+1}} \mu_i^{\frac{\alpha-1}{\alpha+1}} \nu_i^{\frac{1}{\alpha+1}}}{\Xi} \\ &= -\varepsilon K, \end{aligned} \tag{4.22}$$

where

$$K = K(\alpha, \boldsymbol{\kappa}, \boldsymbol{\mu}, \boldsymbol{\nu}) \triangleq \frac{\max_{i\in\mathcal{I}} \kappa_i^{\frac{1}{\alpha+1}} \mu_i^{\frac{\alpha-1}{\alpha+1}} \nu_i^{\frac{1}{\alpha+1}}}{\Xi} \tag{4.23}$$

is a positive load-dependent constant. The second inequality follows from the fact that for any vector $\mathbf{x}$, and for any $\alpha > 0$, $\|\mathbf{x}\|_{\alpha+1} \leq \|\mathbf{x}\|_\alpha$. The second to last equality follows from the definition of the w_i (cf. Equation (4.15)).

For the second term in (4.20), we wish to show that if $L(\mathbf{m})$ is sufficiently large, then

$$\frac{1}{2}\mathbf{n}^T \frac{\nabla^2 F(\mathbf{m}+\theta\mathbf{n})}{L^\alpha(\mathbf{m}+\theta\mathbf{n})}\mathbf{n} \leq \frac{\varepsilon}{2}K.$$

Note that with probability 1, either $\mathbf{n} = \mathbf{0}$ or $\mathbf{n} = \pm\mathbf{e}_i$ for some $i \in \mathcal{I}$. Thus

$$\begin{aligned}
\frac{1}{2}\mathbf{n}^T \frac{\nabla^2 F(\mathbf{m}+\theta\mathbf{n})}{L^\alpha(\mathbf{m}+\theta\mathbf{n})}\mathbf{n} &\leq \frac{1}{2}\frac{\max_{i\in\mathcal{I}} [\nabla^2 F(\mathbf{m}+\theta\mathbf{n})]_{ii}}{L^\alpha(\mathbf{m}+\theta\mathbf{n})} \\
&= \frac{\alpha}{2}\frac{\max_{i\in\mathcal{I}} w_i(m_i+\theta n_i)^{\alpha-1}}{\left[\sum_{i\in\mathcal{I}} w_i(m_i+\theta n_i)^{\alpha+1}\right]^{\frac{\alpha}{\alpha+1}}} \\
&\leq \frac{\alpha}{2}\frac{\max_{i\in\mathcal{I}} w_i(m_i+\theta n_i)^{\alpha-1}}{w_{i_0}^{\frac{\alpha}{\alpha+1}}(m_{i_0}+\theta n_{i_0})^{\alpha}} \\
&\leq \frac{\alpha}{2} w_{i_0}^{\frac{1}{\alpha+1}}(m_{i_0}+\theta n_{i_0})^{-1} \\
&\leq \frac{\alpha}{2}\max_{i\in\mathcal{I}} w_i^{\frac{1}{\alpha+1}}(m_{i_0}+\theta n_{i_0})^{-1},
\end{aligned}$$

where $i_0 \in \mathcal{I}$ is such that $w_{i_0}(m_{i_0}+\theta n_{i_0})^{\alpha-1} = \max_{i\in\mathcal{I}} w_i(m_i+\theta n_i)^{\alpha-1}$.

Now note that

$$\frac{\alpha}{2}\max_{i\in\mathcal{I}} w_i^{\frac{1}{\alpha+1}}(m_{i_0}+\theta n_{i_0})^{-1} \leq \frac{\varepsilon}{2}K$$

(where K is defined in (4.23)) if and only if

$$m_{i_0}+\theta n_{i_0} \geq \frac{\alpha \max_{i\in\mathcal{I}} w_i^{\frac{1}{\alpha+1}}}{K}\cdot\frac{1}{\varepsilon},$$

which holds if $L(\mathbf{m}) \geq K'/\varepsilon$ for some appropriately defined load-dependent constant K'. Thus, if $L(\mathbf{m}) \geq K'/\varepsilon$, then

$$\frac{1}{2}\mathbf{n}^T \frac{\nabla^2 F(\mathbf{m}+\theta\mathbf{n})}{L^\alpha(\mathbf{m}+\theta\mathbf{n})}\mathbf{n} \leq \frac{\varepsilon}{2}K. \tag{4.24}$$

By adding (4.22) and (4.24), we conclude that

$$\mathbb{E}\left[L(\mathbf{m}+\mathbf{n}) - L(\mathbf{m}) \mid \mathbf{m}\right] \leq -\frac{\varepsilon}{2}K,$$

when $L(\mathbf{m}) \geq K'/\varepsilon$.

Proof of Theorem 4.5.3: $\alpha \in (0,1)$. We now consider the case $\alpha \in (0,1)$. The proof in this section is similar to that for the case $\alpha \geq 1$. We invoke Proposition 3.3.1 to write the drift term as a sum of terms, which we bound separately. As in the previous section, we drop the subscript α from L_α, F_α, H_α, and h_α, and write instead L, F, H, and h, respectively. Note that to use Proposition 3.3.1, we need L to be twice continuously differentiable. Indeed, by Lemma 4.5.2 (i), h is continuously differentiable, so its antiderivative H is twice continuously differentiable, and so is L. Thus, by the second order mean value theorem, we obtain an equation similar to Equation (4.20):

$$\begin{aligned}
L(\mathbf{m}+\mathbf{n}) - L(\mathbf{m}) &= \mathbf{n}^T\nabla L(\mathbf{m}) + \frac{1}{2}\mathbf{n}^T\nabla^2 L(\mathbf{m}+\theta\mathbf{n})\mathbf{n} && (4.25)\\
&\leq \frac{\sum_{i\in I} n_i w_i h(m_i)}{L^\alpha(\mathbf{m})} + \frac{1}{2}\frac{\sum_{i\in\mathcal{I}} n_i^2 w_i h'(m_i+\theta n_i)}{L^\alpha(\mathbf{m}+\theta\mathbf{n})} && (4.26)\\
&\leq \frac{\sum_{i\in I} n_i w_i h(m_i)}{L^\alpha(\mathbf{m})} + \frac{1}{2}\frac{\max_{i\in\mathcal{I}} w_i h'(m_i+\theta n_i)}{L^\alpha(\mathbf{m}+\theta\mathbf{n})} && (4.27)
\end{aligned}$$

for some constant $\theta \in [0,1]$, and where, as before, $\tilde{\mathbf{M}}(\tau) = \mathbf{m}$ and $\tilde{\mathbf{M}}(\tau+1) = \mathbf{m}+\mathbf{n}$, and the last inequality follows from the fact that with probability 1, either $\mathbf{n} = \mathbf{0}$, or $\mathbf{n} = \pm\mathbf{e}_i$, for some $i \in \mathcal{I}$, and that h' is nonnegative.

We now bound the two terms in (4.27) separately. Let us first concentrate on the term

$$\frac{\sum_{i\in I} n_i w_i h(m_i)}{L^\alpha(\mathbf{m})}.$$

By Lemma 4.5.2 (iii),

$$\sum_{i\in I} n_i w_i h(m_i) \leq \sum_{i\in I} n_i w_i (m_i^\alpha + 1) \leq \sum_{i\in I} n_i w_i m_i^\alpha + \sum_{i\in I} n_i w_i,$$

so

$$\frac{\sum_{i\in I} n_i w_i h(m_i)}{L^\alpha(\mathbf{m})} \leq \frac{\sum_{i\in I} n_i w_i m_i^\alpha}{L^\alpha(\mathbf{m})} + \frac{\sum_{i\in I} n_i w_i}{L^\alpha(\mathbf{m})}.$$

First consider the term $\frac{\sum_{i\in I} n_i w_i m_i^\alpha}{L^\alpha(\mathbf{m})}$. Note that $\sum_{i\in I} n_i w_i m_i^\alpha = \mathbf{n}^T \nabla F(\mathbf{m})$. We also recall from the proof of Lemma 4.4.3 that

$$\mathbb{E}\left[\mathbf{n}^T \nabla F(\mathbf{m}) \mid \mathbf{m}\right] = \frac{\langle \nabla F(\mathbf{m}), \boldsymbol{\nu} - \boldsymbol{\mu}\boldsymbol{\phi}(\mathbf{m})\rangle}{\Xi} \leq -\varepsilon \frac{\langle \nabla F(\mathbf{m}), \boldsymbol{\nu}\rangle}{\Xi}.$$

We then proceed along the same lines as in the case $\alpha \geq 1$, and obtain that if $L(\mathbf{m}) \geq K_2/\varepsilon$ for some positive load-dependent constant K_2, then

$$\begin{aligned}
\mathbb{E}\left[\frac{\sum_{i\in I} n_i w_i m_i^\alpha}{L^\alpha(\mathbf{m})} \;\middle|\; \mathbf{m}\right] &\leq -\frac{3}{4}\varepsilon \frac{\max_{i\in\mathcal{I}} w_i^{\frac{1}{\alpha+1}} \nu_i}{\Xi} \\
&= -\frac{3}{4}\varepsilon \frac{\max_{i\in\mathcal{I}} \kappa^{\frac{1}{\alpha+1}} \mu_i^{\frac{\alpha-1}{\alpha+1}} \nu_i^{\frac{1}{\alpha+1}}}{\Xi} \\
&= -\frac{3}{4}\varepsilon K, \qquad (4.28)
\end{aligned}$$

Here as in the proof for the case $\alpha \geq 1$, $K = K(\alpha, \boldsymbol{\kappa}, \boldsymbol{\mu}, \boldsymbol{\nu}) \triangleq \frac{\max_{i\in\mathcal{I}} \kappa^{\frac{1}{\alpha+1}} \mu_i^{\frac{\alpha-1}{\alpha+1}} \nu_i^{\frac{1}{\alpha+1}}}{\sum_{i\in\mathcal{I}} \nu_i}$ is a positive load-dependent constant.

Now consider the term $\frac{\sum_{i\in I} n_i w_i}{L^\alpha(\mathbf{m})}$. With probability 1, either $\mathbf{n} = \mathbf{0}$ or $\mathbf{n} = \pm\mathbf{e}_i$ for some $i \in \mathcal{I}$, and therefore $\sum_{i\in\mathcal{I}} n_i w_i \leq \max_{i\in\mathcal{I}} w_i$. Thus,

$$\mathbb{E}\left[\frac{\sum_{i\in I} n_i w_i h(m_i)}{L^\alpha(\mathbf{m})} \;\middle|\; \mathbf{m}\right] \leq -\frac{3}{4}\varepsilon K + \frac{\max_{i\in\mathcal{I}} w_i}{L^\alpha(\mathbf{m})}.$$

For the second term in (4.27), note that with $\alpha \in (0,1)$, Lemma 4.5.2(iv) implies that $h' \leq 2$, and therefore,

$$\frac{1}{2}\frac{\max_{i\in\mathcal{I}} w_i h'(m_i + \theta n_i)}{L^\alpha(\mathbf{m}+\theta\mathbf{n})} \leq \frac{\max_{i\in\mathcal{I}} w_i}{L^\alpha(\mathbf{m}+\theta\mathbf{n})}.$$

Note that $L^\alpha(\mathbf{m}+\theta\mathbf{n})$ and $L^\alpha(\mathbf{m})$ differ only by a load-dependent constant, since with probability 1, either $\mathbf{n} = \mathbf{0}$ or $\mathbf{n} = \pm\mathbf{e}_i$ for some $i \in \mathcal{I}$. Thus, if $L^\alpha(\mathbf{m}) \geq K_3/\varepsilon$

for some positive load-dependent constant K_3, then

$$\frac{\max_{i\in\mathcal{I}} w_i}{L^\alpha(\mathbf{m})} + \frac{\max_{i\in\mathcal{I}} w_i}{L^\alpha(\mathbf{m}+\theta\mathbf{n})} \leq \frac{1}{4}\varepsilon K. \tag{4.29}$$

Putting (4.28) and (4.29) together, we get that if $L(\mathbf{m}) \geq K'/\min\{\varepsilon^{1/\alpha}, \varepsilon\}$, where $K' = \max\{K_3^{1/\alpha}, K_2\}$, then

$$\mathbb{E}\left[L(\mathbf{m}+\mathbf{n}) - L(\mathbf{m}) \mid \mathbf{m}\right] \leq -\frac{\varepsilon}{2}K.$$

4.5.2 Exponential Tail Bound under α-Fair Policies

In this section, we derive an exponential upper bound on the tail probability of the stationary distribution of the flow sizes, under an α-fair policy with $\alpha > 0$. The following theorem, an exact copy of Theorem 3.4.4, and a modification of Theorem 1 from [4] will be used to derive the exponential upper bound.

Theorem 4.5.4 *Let $\mathbf{X}(\cdot)$ be an irreducible and aperiodic discrete-time Markov chain with a countable state space $\mathcal{X}$. Suppose that there exists a Lyapunov function $f : \mathcal{X} \to \mathbb{R}_+$ with the following properties:*

*(a) f has **bounded increments**: there exists $\xi > 0$ such that for all τ, we have*

$$|f(\mathbf{X}(\tau+1)) - f(\mathbf{X}(\tau))| \leq \xi, \quad \textit{almost surely};$$

*(b) **Negative drift**: there exist $B > 0$ and $\gamma > 0$ such that whenever $f(\mathbf{X}(\tau)) > B$,*

$$\mathbb{E}[f(\mathbf{X}(\tau+1)) - f(\mathbf{X}(\tau)) \mid \mathbf{X}(\tau)] \leq -\gamma.$$

Then, a stationary probability distribution π exists, and we have an exponential upper bound on the tail probability of f under π: for any $\ell \in \mathbb{Z}_+$,

$$\mathbb{P}_\pi(f(\mathbf{X}) > B + 2\xi\ell) \leq \left(\frac{\xi}{\xi+\gamma}\right)^{\ell+1}. \tag{4.30}$$

In particular, in steady state, all moments of f are finite, i.e., for every $k \in \mathbb{N}$,

$$\mathbb{E}_\pi[f^k(\mathbf{X})] < \infty.$$

Theorem 4.5.4 is identical to Theorem 1 in [4] except that [4] imposed the additional condition $\mathbb{E}_\pi[f(\mathbf{X})] < \infty$. However, the latter condition is redundant. Indeed, using Foster-Lyapunov criteria (see [18], for example), conditions (a) and (b) in Theorem 4.5.4 imply that the Markov chain $\mathbf{X}$ has a unique stationary distribution π. Furthermore, Theorem 2.3 in [27] establishes that under conditions (a) and (b), all moments of $f(\mathbf{X})$ are finite in steady state. We note that Theorem 2.3 in [27] and Theorem 1 of [4] provide the same qualitative information (exponential tail bounds for $f(\mathbf{X})$). However, [4] contains the more precise bound (4.30), which we will use to prove Theorem 4.6.6 in Section 4.6.

Proof of Theorem 4.3.2. The finiteness of the moments follows immediately from the bound in (4.30), so we only prove the exponential bound (4.30). We apply Theorem 4.5.4 to the Lyapunov function L_α and the uniformized chain $\tilde{\mathbf{M}}(\cdot)$. Again, denote the stationary distribution of $\tilde{\mathbf{M}}(\cdot)$ by $\boldsymbol{\pi}$, and note that this is also the unique stationary distribution of $\mathbf{M}(\cdot)$. The proof consists of verifying conditions (a) and (b).

(a) **Bounded Increments.** We wish to show that with probability 1, there exists ξ such that

$$|L_\alpha(\tilde{\mathbf{M}}(\tau+1)) - L_\alpha(\tilde{\mathbf{M}}(\tau))| \leq \xi.$$

As usual, write $\mathbf{m} = \tilde{\mathbf{M}}(\tau)$ and $\mathbf{m} + \mathbf{n} = \tilde{\mathbf{M}}(\tau+1)$, then $\mathbf{n} = \mathbf{0}$ or $\mathbf{n} = \pm\mathbf{e}_i$ for some $i \in \mathcal{I}$ with probability 1. For $\alpha \geq 1$,

$$L_\alpha(\mathbf{m}) = \left[\sum_{i\in\mathcal{I}} w_i m_i^{\alpha+1}\right]^{\frac{1}{\alpha+1}},$$

and for $\alpha \in (0,1)$, by Lemma 4.5.2 (iii'), we have

$$\sum_{i\in\mathcal{I}} w_i m_i^{\alpha+1} - 2\sum_{i\in\mathcal{I}} w_i \le (\alpha+1)\sum_{i\in\mathcal{I}} w_i H_\alpha(m_i) \le \sum_{i\in\mathcal{I}} w_i m_i^{\alpha+1} + 2\sum_{i\in\mathcal{I}} w_i.$$

In general, for $r, s \ge 0$ and $\beta \in [0,1]$,

$$(r+s)^\beta \le r^\beta + s^\beta. \tag{4.31}$$

Thus, by inequality (4.31),

$$\left[\sum_{i\in\mathcal{I}} w_i m_i^{\alpha+1}\right]^{\frac{1}{\alpha+1}} - \left[2\sum_{i\in\mathcal{I}} w_i\right]^{\frac{1}{\alpha+1}} \le L_\alpha(\mathbf{m}) \le \left[\sum_{i\in\mathcal{I}} w_i m_i^{\alpha+1}\right]^{\frac{1}{\alpha+1}} + \left[2\sum_{i\in\mathcal{I}} w_i\right]^{\frac{1}{\alpha+1}}.$$

Hence, for any $\alpha > 0$,

$$\begin{aligned}
|L_\alpha(\mathbf{m}+\mathbf{n}) - L_\alpha(\mathbf{m})| &\le \left| \left[\sum_{i\in\mathcal{I}} w_i (m_i+n_i)^{\alpha+1}\right]^{\frac{1}{\alpha+1}} - \left[\sum_{i\in\mathcal{I}} w_i m_i^{\alpha+1}\right]^{\frac{1}{\alpha+1}} \right| \\
&\quad + 2\left[2\sum_{i\in\mathcal{I}} w_i\right]^{\frac{1}{\alpha+1}} \\
&\le \left[\sum_{i\in\mathcal{I}} w_i |n_i|^{\alpha+1}\right]^{\frac{1}{\alpha+1}} + 2\left[2\sum_{i\in\mathcal{I}} w_i\right]^{\frac{1}{\alpha+1}} \\
&\le \max_{i\in\mathcal{I}} w_i^{\frac{1}{\alpha+1}} + 2\left[2\sum_{i\in\mathcal{I}} w_i\right]^{\frac{1}{\alpha+1}},
\end{aligned}$$

where the second last inequality follows from the triangle inequality. Thus we can take $\xi = \max_{i\in\mathcal{I}} w_i^{\frac{1}{\alpha+1}} + 2\left[2\sum_{i\in\mathcal{I}} w_i\right]^{\frac{1}{\alpha+1}}$, which is a load-dependent constant.

(b) **Negative Drift.** The negative drift condition is established in Theorem 4.5.3, with $\gamma = \varepsilon K$, for some positive load-dependent constant K.

Note that we have verified conditions (a) and (b) for the Lyapunov function L_α. To show the actual exponential probability tail bound for $\|\mathbf{M}\|_\infty$, note that $L_\alpha(\mathbf{M}) \ge$

$K''\|\mathbf{M}\|_\infty$, for some load-dependent constant K''. By suitably redefining the constants B, ξ, and K, the same form of exponential probability tail bound is established for $\|\mathbf{M}\|_\infty$.

4.6 An Important Application: Interchange of Limits ($\alpha = 1$)

In this section, we assume throughout that $\alpha = 1$ (the proportionally-fair policy), and establish the validity of the heavy-traffic approximation for networks in steady state. We first provide the necessary preliminaries to state our main theorem, Theorem 4.6.6. Further definitions and background will be provided in Section 4.6.2, along with the proof of Theorem 4.6.6. All definitions and background stated in this section are taken from [36] and [32].

4.6.1 Preliminaries

We give a preview of the preliminaries that we will introduce before stating Theorem 4.6.6. The goal of this subsection is to provide just enough background to be able to state Theorem 4.6.5, the diffusion approximation result from [32]. To do this, we need a precise description of the process obtained in the limit, under the diffusion scaling. This limiting process is a diffusion process, called Semimartingale Reflecting Brownian Motion (SRBM) (Definition 4.6.3), with support on a polyhedral cone. This polyhedral cone is defined through the concept of an invariant manifold (Definition 4.6.2).

As in Section 4.4.3, we consider a sequence of networks indexed by r, where r is to be thought of as increasing to infinity along a sequence. The incidence matrix $\mathbf{R}$, the capacity vector $\mathbf{C}$, and the weight vector $\boldsymbol{\kappa}$ do not vary with r. Recall the *heavy-traffic* condition — Assumption 4.4.6, and the definitions of the workload $\mathbf{w}$, the workload process $\mathbf{W}$, and the lifting map Δ from Definition 4.4.7. We carry the notation from Section 4.4.3, so that $\boldsymbol{\theta} > \mathbf{0}$, and $\boldsymbol{\nu}^r \to \boldsymbol{\nu} > \mathbf{0}$, $\boldsymbol{\mu}^r \to \boldsymbol{\mu} > \mathbf{0}$ and

$r(\mathbf{C} - \mathbf{R}\boldsymbol{\lambda}^r) \to \boldsymbol{\theta}$ as $r \to \infty$. Recall that $\mathbf{R}\boldsymbol{\lambda} = \mathbf{C}$. Let $\hat{\mathbf{M}}^r$ and $\hat{\mathbf{W}}^r$ be as in (4.13).

The continuity of the lifting map Δ will be useful in the sequel.

Proposition 4.6.1 (Proposition 4.1 in [32]) *The function $\Delta : \mathbb{R}_+^J \to \mathbb{R}_+^N$ is continuous. Furthermore, for each $\mathbf{w} \in \mathbb{R}_+^J$ and $c > 0$,*

$$\Delta(c\mathbf{w}) = c\Delta(\mathbf{w}). \tag{4.32}$$

Definition 4.6.2 (Invariant manifold) *A state $\mathbf{m} \in \mathbb{R}_+^N$ is called* invariant *if $\mathbf{m} = \Delta(\mathbf{w})$, where $\mathbf{w} = \mathbf{R}\mathbf{E}^{-1}\mathbf{m}$ is the workload, and Δ the lifting map defined in Definition 4.4.7. The set of all invariant states is called the* invariant manifold, *and we denote it by $\mathscr{M}$. We also define the* workload cone *$\mathscr{W}$ by $\mathscr{W} = \mathbf{R}\mathbf{E}^{-1}\mathscr{M}$, where $\mathbf{E} = \mathrm{diag}(\boldsymbol{\mu})$ is as defined in Definition 4.4.7.*

The invariant manifold $\mathscr{M}$ is a polyhedral cone and admits an explicit characterization: we can write it as

$$\mathscr{M} = \left\{ \mathbf{m} \in \mathbb{R}_+^N : m_i = \frac{\lambda_i (\mathbf{y}^T \mathbf{R})_i}{\kappa_i} \text{ for all } i \in \mathcal{I}, \text{ for some } \mathbf{y} \in \mathbb{R}_+^J \right\}.$$

Denote the j-th face of $\mathscr{M}$ by $\mathscr{M}^j$, which can be written as

$$\mathscr{M}^j \triangleq \left\{ \mathbf{m} \in \mathbb{R}_+^N : \; m_i = \frac{\lambda_i (\mathbf{y}^T \mathbf{R})_i}{\kappa_i} \text{ for all } i \in \mathcal{I}, \right.$$
$$\left. \text{for some } \mathbf{y} \in \mathbb{R}_+^J \text{ satisfying } y_j = 0 \right\}.$$

Similarly, denote the j-th face of $\mathscr{W}$ by $\mathscr{W}^j$, which can be written as

$$\mathscr{W}^j \triangleq \mathbf{R}\mathbf{E}^{-1}\mathscr{M}^j.$$

Semimartingale Reflecting Brownian Motion (SRBM).

Definition 4.6.3 *Define the covariance matrix*

$$\boldsymbol{\Gamma} = 2\mathbf{R}\mathbf{E}^{-1} diag(\boldsymbol{\nu})\mathbf{E}^{-1}\mathbf{R}^T.$$

An SRBM *that lives in the cone* $\mathcal{W}$*, has direction of reflection* $\mathbf{e}_j$ *(the* j*th unit vector) on the boundary* $\mathcal{W}^j$ *for each* $j \in \mathcal{J}$*, has drift* $-\boldsymbol{\theta}$ *and covariance* $\boldsymbol{\Gamma}$*, and has initial distribution* $\boldsymbol{\eta}_0$ *on* $\mathcal{W}$ *is an adapted,* J*-dimensional process* $\hat{\mathbf{W}}(\cdot)$ *defined on some filtered probability space* $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$ *such that:*

(i) $\mathbb{P}$*-a.s.,* $\hat{\mathbf{W}}(t) = \hat{\mathbf{W}}(0) + \hat{\mathbf{X}}(t) + \hat{\mathbf{U}}(t)$ *for all* $t \geq 0$*;*

(ii) $\mathbb{P}$*-a.s.,* $\hat{\mathbf{W}}(\cdot)$ *has continuous sample paths,* $\hat{\mathbf{W}}(t) \in \mathcal{W}$ *for all* $t \geq 0$*, and* $\hat{\mathbf{W}}(0)$ *has initial distribution* $\boldsymbol{\eta}_0$*;*

(iii) under $\mathbb{P}$*,* $\hat{\mathbf{X}}(\cdot)$ *is a* J*-dimensional Brownian motion starting at the origin with drift* $-\boldsymbol{\theta}$ *and covariance matrix* $\boldsymbol{\Gamma}$*;*

(iv) for each $j \in \mathcal{J}$*,* $\hat{U}_j(\cdot)$ *is an adapted, one-dimensional process such that* $\mathbb{P}$*-a.s.,*

(a) $\hat{U}_j(0) = 0$*;*

(b) $\hat{U}_j$ *is continuous and non-decreasing;*

(c) $\hat{U}_j(t) = \int_0^t \mathbb{I}_{\{\hat{\mathbf{W}}(s) \in \mathcal{W}^j\}} d\hat{U}_j(s)$ *for all* $t \geq 0$*.*

The process $\hat{\mathbf{W}}(\cdot)$ *is called an SRBM with the data* $(\mathcal{W}, -\boldsymbol{\theta}, \boldsymbol{\Gamma}, \{\mathbf{e}_j : j \in \mathcal{J}\}, \boldsymbol{\eta}_0)$*.*

Diffusion Approximation for $\alpha = 1$.

Assumption 4.6.4 (Local traffic) *For each* $j \in \mathcal{J}$*, there exists at least one* $i \in \mathcal{I}$ *such that* $R_{ji} > 0$ *and* $R_{ki} = 0$ *for all* $k \neq j$*.*

Under the local traffic condition, a diffusion approximation holds.

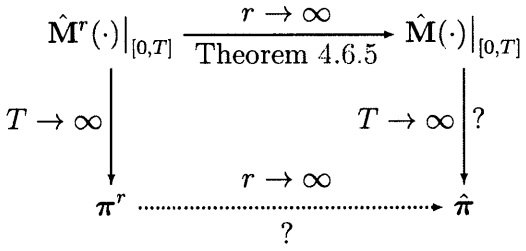
Theorem 4.6.5 (Theorem 5.2 in [32]) *Assume that* $\alpha = 1$ *and that the local traffic condition, Assumption 4.6.4, holds. Suppose that the limit distribution of* $\hat{\mathbf{W}}^r(0)$ *as* $r \to \infty$ *is* $\boldsymbol{\eta}_0$ *(a probability measure on* $\mathcal{W}$*) and that*

$$\|\hat{\mathbf{M}}^r(0) - \Delta(\hat{\mathbf{W}}^r(0))\|_\infty \to 0, \quad \text{in probability, as } r \to \infty. \tag{4.33}$$

Then, the distribution of $(\hat{\mathbf{W}}^r(\cdot), \hat{\mathbf{M}}^r(\cdot))$ *converges weakly (on compact time intervals) as* $r \to \infty$ *to a continuous process* $(\hat{\mathbf{W}}(\cdot), \hat{\mathbf{M}}(\cdot))$*, where* $\hat{\mathbf{W}}(\cdot)$ *is an SRBM with data* $(\mathcal{W}, -\boldsymbol{\theta}, \boldsymbol{\Gamma}, \{\mathbf{e}_j, j \in \mathcal{J}\}, \boldsymbol{\eta}_0)$ *and* $\hat{\mathbf{M}}(t) = \Delta(\hat{\mathbf{W}}(t))$ *for all* t*.*

4.6.2 Interchange of Limits

We now know that for $\alpha = 1$, under the local traffic condition, the diffusion approximation holds. That is, the scaled process $(\hat{\mathbf{W}}^r(\cdot), \hat{\mathbf{M}}^r(\cdot))$ converges in distribution to $(\hat{\mathbf{W}}(\cdot), \hat{\mathbf{M}}(\cdot))$, with $\hat{\mathbf{W}}(\cdot)$ being an SRBM. For any r, the scaled processes $\hat{\mathbf{M}}^r(\cdot)$ also have stationary distributions $\boldsymbol{\pi}^r$, since they are all positive recurrent. These results can be summarized in the diagram that follows.



As can be seen from the diagram, two natural questions to ask are:

1. Does the diffusion process $\hat{\mathbf{M}}(\cdot)$ have a stationary probability distribution, $\hat{\boldsymbol{\pi}}$?

2. If $\hat{\boldsymbol{\pi}}$ exists and is unique, do the distributions $\boldsymbol{\pi}^r$ converge to $\hat{\boldsymbol{\pi}}$?

Our contribution here is a positive answer to question 2. More specifically, if $\hat{\mathbf{M}}(\cdot)$ has a unique stationary probability distribution $\hat{\boldsymbol{\pi}}$, then $\boldsymbol{\pi}^r$ converges in distribution to $\hat{\boldsymbol{\pi}}$.

Theorem 4.6.6 *Suppose that $\alpha = 1$ and that the local traffic condition, Assumption 4.6.4, holds. Suppose further that $\hat{\mathbf{M}}(\cdot)$ has a unique stationary probability distribution $\hat{\boldsymbol{\pi}}$. For each r, let $\boldsymbol{\pi}^r$ be the unique stationary probability distribution of $\hat{\mathbf{M}}^r$. Then,*

$$\boldsymbol{\pi}^r \to \hat{\boldsymbol{\pi}}, \text{ in distribution, as } r \to \infty.$$

The line of proof of Theorem 4.6.6 is fairly standard. We first establish tightness of the set of distributions $\{\boldsymbol{\pi}^r\}$ in Lemma 4.6.7. Letting the processes $\hat{\mathbf{M}}^r(\cdot)$ be initially distributed as $\{\boldsymbol{\pi}^r\}$, we translate this tightness condition into an initial condition similar to (4.33), in Lemma 4.6.8. We then apply Theorem 4.6.5 to deduce the

convergence of the processes $\hat{\mathbf{M}}^r(\cdot)$, which by stationarity, leads to the convergence of the distributions $\boldsymbol{\pi}^r$. We state Lemmas 4.6.7 and 4.6.8 below, and defer their proofs to Appendix A.

We now remark on the validity of Theorem 4.6.6 under more general conditions. Both lemmas 4.6.7 and 4.6.8 hold when $\alpha \geq 1$. Thus, if the diffusion approximation holds when $\alpha \geq 1$, then Theorem 4.6.6 holds as well. However, Theorem 4.6.6 uses Theorem 4.6.5, the diffusion approximation when $\alpha = 1$, and hence we require the condition $\alpha = 1$ and the local-traffic condition.

Lemma 4.6.7 *Suppose that $\alpha = 1$. The set of probability distributions $\{\boldsymbol{\pi}^r\}$ is tight.*

Lemma 4.6.8 *Consider the stationary probability distributions $\boldsymbol{\pi}^r$ of $\hat{\mathbf{M}}^r(\cdot)$, and let $\{\boldsymbol{\pi}^{r_k}\}$ be any convergent subsequence of $\{\boldsymbol{\pi}^r\}$. Let $\hat{\mathbf{M}}^r(0)$ be distributed as $\boldsymbol{\pi}^r$ for each r. Then there exists a subsequence r_ℓ of r_k such that*

$$\left\| \hat{\mathbf{M}}^{r_\ell}(0) - \Delta\left(\hat{\mathbf{W}}^{r_\ell}(0)\right) \right\|_\infty \to 0 \tag{4.34}$$

in probability as $\ell \to \infty$, i.e., such that condition (4.33) holds for the subsequence $\{(\hat{\mathbf{W}}^{r_\ell}(\cdot), \hat{\mathbf{M}}^{r_\ell}(\cdot))\}$.

Proof of Theorem 4.6.6. Since $\{\boldsymbol{\pi}^r\}$ is tight, by Lemma 4.6.7, Prohorov's theorem implies that $\{\boldsymbol{\pi}^r\}$ is relatively compact in the weak topology. Let $\{\boldsymbol{\pi}^{r_k}\}$ be a convergent subsequence of the set of probability distributions $\{\boldsymbol{\pi}^r\}$, and suppose that $\boldsymbol{\pi}^{r_k} \to \boldsymbol{\pi}$ as $k \to \infty$, in distribution.

Let $\hat{\mathbf{M}}^r(0)$ be distributed as $\boldsymbol{\pi}^r$ for each r. Then by Lemma 4.6.8, there exists a subsequence r_ℓ of r_k such that

$$\left\| \hat{\mathbf{M}}^{r_\ell}(0) - \Delta\left(\hat{\mathbf{W}}^{r_\ell}(0)\right) \right\|_\infty \to 0$$

in probability as $\ell \to \infty$. Denote the distribution of $\hat{\mathbf{W}}^r(0)$ by $\boldsymbol{\eta}^r$. Since $\boldsymbol{\pi}^{r_k} \to \boldsymbol{\pi}$ as $k \to \infty$, $\boldsymbol{\pi}^{r_\ell} \to \boldsymbol{\pi}$ as $\ell \to \infty$ as well, and $\boldsymbol{\eta}^{r_\ell} \to \boldsymbol{\eta}$ as $\ell \to \infty$, for some probability distribution $\boldsymbol{\eta}$.

We now wish to apply Theorem 4.6.5 to the sequence $\{\hat{\mathbf{M}}^{r_\ell}(\cdot)\}$. The only condition that needs to be verified is that $\boldsymbol{\eta}$ has support on $\mathscr{W}$. This can be argued as follows. Let $\hat{\mathbf{M}}(0)$ have distribution $\boldsymbol{\pi}$, and let $\hat{\mathbf{W}}(0) = \mathbf{R}\mathbf{E}^{-1}\hat{\mathbf{M}}(0)$ be the corresponding workload. Then $\hat{\mathbf{W}}^{r_\ell}(0) \to \hat{\mathbf{W}}(0)$ in distribution as $r \to \infty$, and $\hat{\mathbf{W}}(0)$ has distribution $\boldsymbol{\eta}$. The lifting map Δ is continuous by Proposition 4.6.1, so $\Delta\left(\hat{\mathbf{W}}^{r_\ell}(0)\right) \to \Delta\left(\hat{\mathbf{W}}(0)\right)$ in distribution as $r \to \infty$. This convergence, together with (4.34), and the fact that $\hat{\mathbf{M}}^{r_\ell}(0) \to \hat{\mathbf{M}}(0)$ in distribution, implies that $\hat{\mathbf{M}}(0)$ and $\Delta\left(\hat{\mathbf{W}}(0)\right)$ are identically distributed. Now $\Delta\left(\hat{\mathbf{W}}(0)\right)$ has support on $\mathscr{M}$, so $\hat{\mathbf{M}}(0)$ is supported on $\mathscr{M}$ as well, and so $\hat{\mathbf{W}}(0)$, hence $\boldsymbol{\eta}$, is supported on $\mathscr{W}$.

By Theorem 4.6.5, $(\hat{\mathbf{W}}^{r_\ell}(\cdot), \hat{\mathbf{M}}^{r_\ell}(\cdot))$ converges in distribution to a continuous process $(\hat{\mathbf{W}}(\cdot), \hat{\mathbf{M}}(\cdot))$. Suppose that $\hat{\mathbf{W}}(\cdot)$ and $\hat{\mathbf{M}}(\cdot)$ have unique stationary distributions $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\pi}}$, respectively. The processes $(\hat{\mathbf{W}}^{r_\ell}(\cdot), \hat{\mathbf{M}}^{r_\ell}(\cdot))$ are stationary, so $(\hat{\mathbf{W}}(\cdot), \hat{\mathbf{M}}(\cdot))$ is stationary as well. Therefore, $\hat{\mathbf{W}}(0)$ and $\hat{\mathbf{M}}(0)$ are distributed as $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\pi}}$, respectively. Since $(\hat{\mathbf{W}}^{r_\ell}(0), \hat{\mathbf{M}}^{r_\ell}(0)) \to (\hat{\mathbf{W}}(0), \hat{\mathbf{M}}(0))$ in distribution, we have that $\boldsymbol{\eta}^{r_\ell} \to \hat{\boldsymbol{\eta}}$ and $\boldsymbol{\pi}^{r_\ell} \to \hat{\boldsymbol{\pi}}$ weakly as $\ell \to \infty$. This shows that $\boldsymbol{\pi} = \hat{\boldsymbol{\pi}}$ and $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}$. Since $\{\boldsymbol{\pi}^{r_k}\}$ is an arbitrary convergent subsequence $\hat{\boldsymbol{\pi}}$ is the unique weak limit point of $\{\boldsymbol{\pi}^r\}$, and this shows that $\boldsymbol{\pi}^r \to \hat{\boldsymbol{\pi}}$ in distribution. □

For Theorem 4.6.6 to apply, we need to verify that $\hat{\mathbf{M}}(\cdot)$ (or equivalently, $\hat{\mathbf{W}}(\cdot)$) has a unique stationary distribution. The following theorem states that when $\kappa_i = 1$ for all $i \in \mathcal{I}$, this condition holds; more specifically, the SRBM $\hat{\mathbf{W}}(\cdot)$ has a unique stationary distribution, which turns out to have a product form.

Theorem 4.6.9 (Theorem 5.3 in [32]) *Suppose that $\alpha = 1$ and $\kappa_i = 1$ for all $i \in \mathcal{I}$. Let $\hat{\boldsymbol{\eta}}$ be the measure on $\mathscr{W}$ that is absolutely continuous with respect to Lebesgue measure with density given by*

$$p(\mathbf{w}) = \exp(\langle \mathbf{v}, \mathbf{w} \rangle), \quad \mathbf{w} \in \mathscr{W}, \tag{4.35}$$

where

$$\mathbf{v} = -2\boldsymbol{\Gamma}^{-1}\boldsymbol{\theta}. \tag{4.36}$$

The product measure $\hat{\boldsymbol{\eta}}$ is an invariant measure for the SRBM $\hat{\mathbf{W}}$ with state space $\mathcal{W}$, directions of reflection $\{\mathbf{e}_j, j \in \mathcal{J}\}$, drift $-\boldsymbol{\theta}$, and covariance matrix $\boldsymbol{\Gamma}$. After normalization, it defines the unique *stationary distribution for the SRBM.*

Here we remark on the density (4.35). As pointed out in [32], the product form of (4.35) does not imply that the components of the SRBM $\hat{\mathbf{W}}$ are independent in steady state, since the cone $\mathcal{W}$ is in general not an orthant. However, independence holds for a proper linear transformation of $\hat{\mathbf{W}}$.

Corollary 4.6.10 (Corollary 5.1 in [32]) *Suppose that the assumptions of Theorem 4.6.9 hold. Let $\hat{\mathbf{W}}$ be as in Theorem 4.6.9, and $\boldsymbol{\Gamma}$ be the covariance matrix. Then the SRBM $\hat{\mathbf{X}} = 2\boldsymbol{\Gamma}^{-1}\hat{\mathbf{W}}$ of dual variables has a unique stationary distribution, where the jth component $\hat{X}_j$ of $\hat{\mathbf{X}}$ is an independent exponential random variable with mean $1/\theta_j$ in steady state, for each $j \in \mathcal{J}$.*

By Theorems 4.6.6 and 4.6.9, the following corollary is immediate.

Corollary 4.6.11 *Suppose that $\alpha = 1$ and $\kappa_i = 1$ for all $i \in \mathcal{I}$. Suppose further that the local traffic condition, Assumption 4.6.4, holds. Let $\hat{\boldsymbol{\pi}}$ be the unique stationary probability distribution of $\hat{\mathbf{M}}(\cdot)$. For each r, let $\boldsymbol{\pi}^r$ be the unique stationary probability distribution of $\hat{\mathbf{M}}^r$. Then,*

$$\boldsymbol{\pi}^r \to \hat{\boldsymbol{\pi}}, \text{ in distribution, as } r \to \infty.$$

4.7 Proportional Fairness in Input-Queued Switches

To better understand the implication of results in Section 4.6, we carry out a formal calculation of the performance of proportional fairness in input-queued switches. The calculation is only "formal", because the incidence matrix associated with an input-queued switch fails to satisfy the technical conditions required for a diffusion approximation to hold, and earlier results cannot be applied. Building upon this calculation, we state a conjecture on the performance of proportional fairness in input-queued switches. The high-level idea is that the crisp product-form distribution in

Corollary 4.6.10 allows us to explicitly compute various quantities of interest, and, in this section, the expected total number of flows in diffusion scale.

A Formal Calculation. Here we present a formal calculation of the performance of proportional fairness in input-queued switches. We consider a continuous-time analog of the discrete-time input-queued switch, and consider the performance of proportional fairness in this continuous-time model.

Consider a **BN** with the structure of an $n \times n$ input-queued switch. There are n^2 routes, which correspond to the queues, and $2n$ resources, which correspond to the input and output ports, and each route uses exactly one input-port resource, and one output-port resource. For the route that uses input-port resource i and output-port resource j, we use the label (i, j), following the convention for input-queued switches. We can write down the incidence matrix $\mathbf{R}$. In general, $\mathbf{R}$ is a $2n \times n^2$ matrix, and for example, when $n = 3$, we have

$$\mathbf{R} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

In this example, $\mathbf{R}$ is a 6×9 matrix, where the first three rows correspond to input-port resources, and the last three rows correspond to output-port resources. There are 9 routes, which correspond to the columns, and $R_{k\ell} = 1$ iff route ℓ uses resource k; otherwise $R_{k\ell} = 0$.

For every $n \in \mathbb{N}$, $\mathbf{R}$ has rank $2n - 1$. This means that $\mathbf{R}$ is not full-rank, and in particular, does not satisfy the local traffic condition, Assumption 4.6.4.

We recall some further background. Suppose that all flows that arrive to the **BN** have mean size 1, and that the capacity on each of the $2n$ resources is uniformly 1. Let λ_{ij} be the arrival rate of flows on route (i, j), and let $\boldsymbol{\lambda} = (\lambda_{ij})$ be the arrival rate

vector. Then $\mathbf{R}\boldsymbol{\lambda} < \mathbf{1}$ is a necessary and sufficient condition for stability, where $\mathbf{1}$ is the vector of all ones. If $\mathbf{R}\boldsymbol{\lambda} < \mathbf{1}$, then $\boldsymbol{\lambda}$ is called strictly admissible.

We now consider a sequence of networks indexed by r, where r is to be thought of as tending to infinity. Let the rth network have arrival rates $\boldsymbol{\lambda}^r = (\lambda^r_{ij})_{i,j=1}^n$. Suppose that $\boldsymbol{\lambda}^r$ is strictly admissible, for all r. Consider the heavy-traffic condition given by $\boldsymbol{\lambda}^r \to \hat{\boldsymbol{\lambda}}$ as $r \to \infty$, and $r(\mathbf{1} - \mathbf{R}\boldsymbol{\lambda}) \to \mathbf{1}$. Then $\mathbf{R}\hat{\boldsymbol{\lambda}} = \mathbf{1}$, and in particular, $\sum_{\ell=1}^n \hat{\lambda}_{i,\ell} = 1$ for all $i \in \{1, 2, \ldots, n\}$, and $\sum_{\ell=1}^n \hat{\lambda}_{\ell,j} = 1$ for all $j \in \{1, 2, \ldots, n\}$, i.e., all resources are critically loaded.

Let $\mathbf{N}^r$ be the flow vector in the rth network, and $\mathbf{W}^r$ the workload vector. Then, since all flows have mean size 1, $\mathbf{W}^r = \mathbf{R}\mathbf{N}^r$. Let $\hat{\mathbf{N}}^r$ and $\hat{\mathbf{W}}^r$ be the corresponding processes under diffusion scaling (see (4.13)). We can also formally calculate the covariance matrix (it is not invertible)

$$\boldsymbol{\Gamma} = 2\mathbf{R}\mathbf{E}^{-1}\mathrm{diag}(\hat{\boldsymbol{\lambda}})\mathbf{E}^{-1}\mathbf{R}^T = 2\mathbf{R}\mathrm{diag}(\hat{\boldsymbol{\lambda}})\mathbf{R}^T$$

which turns out to be

$$\boldsymbol{\Gamma} = 2\left(\begin{array}{c|c} \mathbf{I} & \hat{\boldsymbol{\lambda}} \\ \hline \hat{\boldsymbol{\lambda}} & \mathbf{I} \end{array}\right), \tag{4.37}$$

where here $\hat{\boldsymbol{\lambda}}$ has entries $\hat{\lambda}_{ij}$.

Now consider the network under the proportionally fair policy. If the diffusion approximation holds, and a product-form stationary distribution exists for the SRBM under the heavy-traffic limit, then the interchange-of-limits result, Corollary 4.6.11 holds. In particular, the stationary distribution of $\hat{\mathbf{W}}^r$ converges to a product-form distribution as $r \to \infty$, where the product-form distribution has the distribution of $\frac{1}{2}\boldsymbol{\Gamma}\mathbf{X}$, with $\mathbf{X}$ having independent components that are all exponentially distributed with mean 1. ($\mathbf{X}$ can be informally thought of as $2\boldsymbol{\Gamma}^{-1}\hat{\mathbf{W}}(\infty)$ in Corollary 4.6.10. This consideration is informal because $\boldsymbol{\Gamma}$ defined in (4.37) is not invertible.) Then, under this product-form distribution, the limiting workload $\hat{\mathbf{W}}$ satisfies

$$\mathbb{E}[\|\hat{\mathbf{W}}\|_1] = \mathbb{E}\left[\left\|\frac{1}{2}\boldsymbol{\Gamma}\mathbf{X}\right\|_1\right] = \left\|\frac{1}{2}\boldsymbol{\Gamma}\mathbb{E}[\mathbf{X}]\right\|_1 = 4n.$$

Inspecting the structure of $\mathbf{R}$, we also have

$$\mathbb{E}[\|\hat{\mathbf{N}}\|_1] = \frac{1}{2}\mathbb{E}[\|\hat{\mathbf{W}}\|_1] = 2n.$$

This suggests that under the proportionally fair policy, and in the heavy-traffic limit, we should have

$$\lim_{r\to\infty} \mathbb{E}_{\boldsymbol{\pi}^r}[\|\hat{\mathbf{N}}^r\|_1] = 2n,$$

where $\boldsymbol{\pi}^r$ is the stationary distribution of $\hat{\mathbf{N}}^r$, for each r.

The Conjecture. We now state the conjectured performance of proportional fairness in input-queued switches. Consider a sequence of $n \times n$ input-queued switches operating in discrete time, indexed by r (where r is to be thought of as tending to infinity). For all these networks, let the schedule set be $\mathcal{S}$, the admissible region be $\bar{\mathbf{\Lambda}}$, and the strictly admissible region be $\mathbf{\Lambda}$. Let $\mathbf{Q}^r(\cdot)$ be the queue-size process of the rth system. Let $\boldsymbol{\lambda}^r = (\lambda^r_{ij})_{i,j=1}^n \in \mathbf{\Lambda}$ be the arrival rate vector for the rth network, and suppose that for each $i \in \{1, 2, \ldots, n\}$, $r\left(1 - \sum_{\ell=1}^n \lambda^r_{i,\ell}\right) \to 1$ as $r \to \infty$, and for each $j \in \{1, 2, \ldots, n\}$, $r\left(1 - \sum_{\ell=1}^n \lambda^r_{\ell,j}\right) \to 1$ as $r \to \infty$. Furthermore, suppose that $\boldsymbol{\lambda}^r \to \hat{\boldsymbol{\lambda}}$ as $r \to \infty$. Then, we must have

$$\sum_{\ell=1}^n \hat{\lambda}_{i\ell} = \sum_{\ell=1}^n \hat{\lambda}_{\ell j} = 1.$$

The proportionally fair policy operates in continuous time, and here we need a discrete-time analogue. Toward this end, in time slot τ, we first find a rate vector $\boldsymbol{\sigma}(\tau) = \big(\sigma_{ij}(\tau)\big)$ that satisfies

$$\boldsymbol{\sigma}(\tau) = \arg\max_{\boldsymbol{\sigma}\in\mathbf{\Lambda}} \sum_{i,j=1}^n Q_{ij}(\tau) \log \sigma_{ij}.$$

We can write $\boldsymbol{\sigma}(\tau)$ as a convex combination of schedules in the schedule set $\mathcal{S}$. The discrete-time proportionally fair policy then picks a random schedule in $\mathcal{S}$ according to the convex combination specified by $\boldsymbol{\sigma}(\tau)$. Note that under this policy, for each r,

$\mathbf{Q}^r(\cdot)$ is positive recurrent, and hence a unique stationary distribution $\boldsymbol{\pi}^r$ exists. We propose the following conjecture for this policy.

Conjecture 4.7.1 *Consider the sequence of input-queued switches under the heavy-traffic condition described above. Then under the discrete-time proportionally fair policy,*

$$\limsup_{r\to\infty} \mathbb{E}_{\boldsymbol{\pi}^r}\left[\frac{1}{r}\|\mathbf{Q}^r\|_1\right] \leq 2n.$$

The conjecture implies that proportional fairness is heavy-traffic (near-)optimal in input-queued switches. This follows from the fact that the conjectured upper bound is of order $O(n)$, and an $\Omega(n)$ universal lower bound can be obtained for the same quantity, under any policy. The detailed argument used to derive this lower bound can be found in Lemma 5.3.4, Chapter 5. As a prelude, in the next chapter, we will design a scheduling policy in switched networks, based on a so-called Store-and-Forward allocation policy in bandwidth-sharing networks, which is closely related to proportional fairness (see discussion in Section 5.6, Chapter 5). Also note that the scheduling policy that we propose in the next chapter achieves an $O(n)$ upper bound on the expected steady-state total queue size under the diffusion scale, in an $n \times n$ input-queued switch.

4.8 Discussion

Here we provide some discussion on the results in both Chapter 3 and 4. These results can be viewed from two different perspectives. On the one hand, they provide much new information on the qualitative behavior (e.g., finiteness of the expected queue sizes/number of flows, bounds on steady-state tail probabilities and finite-horizon maximum excursion probabilities, etc.) of the important α-weighted resource allocation policies. On the other hand, at an abstract level, our results highlight the choice and analysis of a suitable Lyapunov function. Even if a network is shown to be stable by using a particular Lyapunov function, different choices and more detailed analysis may lead to more powerful bounds. More concretely, we present a

generic method for deriving full state space collapse from multiplicative state space collapse, and one for deriving steady-state exponential tail bounds. We believe that these methods should extend easily to other settings, for example, to general SPNs operating under so-called Maximum-Pressure-β (MP-β) policies [11, 12].

Chapter 5

Optimal Queue-Size Scaling in SN

In previous chapters, we considered various performance properties of some important resource allocation policies, and derived many new insights. One prominent feature of the performance bounds that we obtained is their dependence on both the load, as well as the network structure, or system size. Starting from this chapter, we consider the problem of queue-size scaling in switched networks. We will be explicitly concerned with the dependence of queue sizes on both the network structure and the load factor. We will particularly be interested in the queue-size behavior when the system size N is large, and the load ρ is close to 1.

The main result of this chapter is a new online scheduling policy, which admits performance bounds with explicit dependence on both the network structure and the load, in general single-hop switched networks. An important consequence of the result is that the policy achieves optimal queue-size scaling in the heavy-traffic regime, i.e., when the load ρ goes to 1, for input-queued switches.

The rest of the chapter is organized as follows. We start with a motivating example in Section 5.1 to illustrate how the network structure can affect system queue sizes. We then provide some preliminaries in Section 5.2. In Section 5.3, we first provide a high-level description of our policy, and then state our main result, Theorem 5.3.1. This is followed by a discussion of the optimality of our policy. Section 5.4 details the necessary background on so-called store-and-forward bandwidth allocation (**SFA**) policy, a key component in the design of our policy. We describe our policy in detail,

and prove Theorem 5.3.1 in Section 5.5. A general discussion of possible future work is provided in Section 5.6.

The prerequisite for reading this chapter is the description of the switched network model in Section 2.2, Chapter 2.

5.1 Motivation

Here we provide a simple example to motivate the study undertaken in this chapter.

Consider a work-conserving M/D/1 queue with a unit-rate server in which unit-sized packets arrive as a Poisson process with rate $\rho \in (0,1)$. Then, the average queue size scales[1] as $1/(1-\rho)$. Such scaling dependence of the average queue size on $1/(1-\rho)$ (or the inverse of the *gap*, $1-\rho$, from the load to the capacity) is a universally observed behavior in a large class of queueing networks. In a switched network, the scaling of the average total queue size ought to also depend on the number of queues, N. For example, consider N parallel M/D/1 queues as described above. Clearly, the total average total queue size will scale as $N/(1-\rho)$. On the other hand, consider a variation where all of these queues pool their resources into a single server that works N times faster. Equivalently, by a time change, let each of the N queues receive packets as an independent Poisson process of rate ρ/N, and let each time a common unit-rate server serve a packet from one of the non-empty queues. Then, the average total queue size scales as $1/(1-\rho)$. Indeed, these are instances of switched networks that differ in their scheduling set $\mathcal{S}$, which leads to different queue-size scalings. Therefore, a natural question is the determination of long-run average queue-size scaling in terms of $\mathcal{S}$ and $(1-\rho)$, where ρ is the effective load. In the context of an n-port input-queued switch with $N = n^2$ queues, the optimal scaling of the long-run average total queue size has been conjectured to be $n/(1-\rho)$, that is, $\sqrt{N}/(1-\rho)$ [51].

[1] In this chapter, by scaling of a quantity we mean its dependence (ignoring universal constants) on $\frac{1}{1-\rho}$ and/or the number of queues, N, as these quantities become large. Of particular interest is the scaling when $\rho \to 1$ and $N \to \infty$, in that order.

5.2 Preliminaries

First recall from Section 2.2 that arrival processes are assumed to be Poisson in this chapter. We also assume that the schedule set $\mathcal{S}$ is *monotone*.

Assumption 5.2.1 (Monotonicity) *If $\mathcal{S}$ contains a schedule, then $\mathcal{S}$ also contains all of its sub-schedules. Formally, for any $\boldsymbol{\sigma} \in \mathcal{S}$, if $\boldsymbol{\sigma}' \in \{0,1\}^N$ and $\boldsymbol{\sigma}' \leq \boldsymbol{\sigma}$ componentwise, then $\boldsymbol{\sigma}' \in \mathcal{S}$.*

Under Assumption 5.2.1, the admissible region $\bar{\boldsymbol{\Lambda}}$ and the convex hull $\langle \mathcal{S} \rangle$ of the schedule set $\mathcal{S}$ coincide. In the sequel, we will often use $\bar{\boldsymbol{\Lambda}}$ and $\langle \mathcal{S} \rangle$ interchangeably, depending on the context.

Given that $\langle \mathcal{S} \rangle$ is a polytope contained in $[0,1]^N$, there exists an integer $J \geq 1$, a matrix $\mathbf{R} \in \mathbb{R}_+^{J \times N}$, and a vector $\mathbf{C} \in \mathbb{R}_+^J$ such that

$$\langle \mathcal{S} \rangle = \Big\{ \mathbf{x} \in [0,1]^N : \mathbf{R}\mathbf{x} \leq \mathbf{C} \Big\}. \tag{5.1}$$

We call J the *rank* of $\langle \mathcal{S} \rangle$ (or $\bar{\boldsymbol{\Lambda}}$) in the representation (5.1). When it is clear from the context, we simply call J the rank of $\langle \mathcal{S} \rangle$ (or $\bar{\boldsymbol{\Lambda}}$). Note that this rank may be different from the rank of the matrix $\mathbf{R}$. Our results will exploit the fact that the rank J may be an order of magnitude smaller than N.

5.3 Main result and Its Implications

Before we state our main result, and describe its implications on optimality, we give a high-level description of the policy that we propose. The switched network (**SN**) of interest will be coupled with an appropriate bandwidth-sharing network (**BN**) that operates in continuous time. Under a particular policy known as the "store-and-forward" allocation (**SFA**), the queue-size vector in **BN** has a product-form stationary distribution. Our policy in the original **SN** effectively emulates, in an online manner, the SFA, so as to approximate the product-form stationary distribution in **BN**. As such, we will call our policy **EMUL** (for emulation) from now on. We will describe

in detail **BN** and **SFA** in Section 5.4, and **EMUL** in Section 5.5.

5.3.1 Main Theorem

Theorem 5.3.1 *Consider a single-hop switched network with scheduling set $\mathcal{S}$, admissible region $\bar{\Lambda} = \{\mathbf{x} \in [0,1]^N : \mathbf{R}\mathbf{x} \leq \mathbf{C}\}$ with rank J, and a strictly admissible arrival rate vector $\boldsymbol{\lambda}$ with $\rho = \rho(\boldsymbol{\lambda}) < 1$. Suppose that the system is empty at time 0. Let $\tilde{\rho}_j = \left(\sum_i R_{ji}\lambda_i\right)/C_j$, $j = 1, 2, \ldots, J$. Then under* **EMUL**, *the Markov chain describing the underlying network is positive recurrent, and the queue-size vector $\mathbf{Q}(\cdot)$ has a unique stationary distribution. With respect to this stationary distribution, the following properties hold:*

1. *The expected total queue size is bounded as*

$$\mathbb{E}\Big[\sum_{i=1}^{N} Q_i\Big] \leq \frac{1}{2}\left(\sum_{j=1}^{J} \frac{\tilde{\rho}_j}{1-\tilde{\rho}_j}\right) + K(N+2), \tag{5.2}$$

   where $K = \max_{\boldsymbol{\sigma}\in\mathcal{S}}\left(\sum_i \sigma_i\right)$.

2. *The distribution of the total queue size has an exponential tail with exponent given by*

$$\lim_{L\to\infty} \frac{1}{L} \log \mathbb{P}\Big(\sum_{i=1}^{N} Q_i \geq L\Big) = -\theta^*, \tag{5.3}$$

   where θ^* is the unique positive *solution of the equation $\rho(e^{\theta} - 1) = \theta$.*

5.3.2 Optimality of EMUL in Input-Queued Switches

This section establishes the optimality of our policy for input-queued switches, both with respect to expected total queue size scaling and tail exponent.

Scaling of Queue Sizes. We start by formalizing what we mean by the optimality of expected queue sizes and of their tail exponents. We consider policies under which there is a well-defined limiting stationary distribution of the queue sizes for all $\boldsymbol{\lambda}$ such

that $\rho(\boldsymbol{\lambda}) < 1$. Note that this class of policies is not empty; indeed, the maximum weight policy and our policy are members of this class. With some abuse of notation, let $\boldsymbol{\pi}$ denote the stationary distribution of the queue-size vector under the policy of interest. We are interested in two quantities:

1. *Expected total queue size.* Let $\bar{Q}$ be the expected total queue size under the stationary distribution $\boldsymbol{\pi}$, defined by

$$\bar{Q} = \mathbb{E}_{\boldsymbol{\pi}}[\sum_i Q_i].$$

   Note that by ergodicity, the time average of the total queue size and the expected total queue size under $\boldsymbol{\pi}$ are the same quantity.

2. *Tail exponent.* Let $\beta_L(Q), \beta_U(Q) \in [-\infty, 0]$ be the lower and upper limits of the tail exponent of the total queue size under $\boldsymbol{\pi}$ (possibly $-\infty$ or 0), respectively, defined by

$$\beta_L(Q) = \liminf_{L\to\infty} \frac{1}{L} \log \mathbb{P}_{\boldsymbol{\pi}}\big(\sum_i Q_i \geq L\big), \tag{5.4}$$

$$\text{and} \quad \beta_U(Q) = \limsup_{L\to\infty} \frac{1}{L} \log \mathbb{P}_{\boldsymbol{\pi}}\big(\sum_i Q_i \geq L\big). \tag{5.5}$$

   If $\beta_L(Q) = \beta_U(Q)$, then we denote this common value by $\beta(Q)$.

We are interested in policies that can achieve minimal $\bar{Q}$ and $\beta(Q)$. For tractability reasons, we focus on the *scaling* of these quantities with respect to $\mathcal{S}$ (equivalently, N) and $\rho(\boldsymbol{\lambda})$, as $1/(1-\rho(\boldsymbol{\lambda}))$ and N increase. Now, for different $\boldsymbol{\lambda}'$ and $\boldsymbol{\lambda}$, it is possible that $\rho(\boldsymbol{\lambda}) = \rho(\boldsymbol{\lambda}')$, but the scaling of $\bar{Q}$, for example, could be wildly different. For this reason, we consider the worst possible dependence on $1/(1-\rho)$ and N among all $\boldsymbol{\lambda}$ with $\rho(\boldsymbol{\lambda}) = \rho$.

Note that we are considering scalings with respect to two quantities ρ and N, and we are interested in two limiting regimes $\rho \to 1$ and $N \to \infty$. The optimality of average queue-size stated here is with respect to the order of limits $\rho \to 1$ and then $N \to \infty$. As noted in [51], taking the limits in different orders could potentially result

in different limiting behaviors of the object of interest, e.g., $\bar{Q}$. For more discussion, see Section 5.6. It should be noted, however, that the optimality of the tail exponent holds for *any* ρ and N.

Optimality of The Tail Exponent. Here we establish the optimality of the tail exponent for input-queued switches under our policy. First, we present a universal lower bound on the tail exponent, under any policy, and for a general single-hop switched network. This lower bound is then specialized to the context of input-queued switches, and compared against the tail exponent under our policy.

Consider any policy under which there exists a well-defined limiting stationary distribution of the queue sizes for all $\boldsymbol{\lambda}$ such that $\rho(\boldsymbol{\lambda}) < 1$. Let $\boldsymbol{\pi}_0$ denote the stationary distribution of queue sizes under this policy. The following lemma establishes a universal lower bound on the tail exponent.

Lemma 5.3.2 *Consider a switched network as described in Theorem 5.3.1, with scheduling set* $\mathcal{S}$ *and admissible region* $\{\mathbf{x} \in [0,1]^N : \mathbf{Rx} \leq \mathbf{C}\}$. *Let* $\boldsymbol{\pi}_0$ *and* $\boldsymbol{\lambda}$ *be as described. For each* j, *let* $\tilde{\rho}_j = \sum_{i=1}^N R_{ji}\lambda_i/C_j$ *be defined as in Theorem 5.3.1. Then under* $\boldsymbol{\pi}_0$,

$$\liminf_{L\to\infty} \frac{1}{L} \log \mathbb{P}_{\boldsymbol{\pi}_0} \left(\sum_i Q_i \geq L \right) \geq - \min_{j=1,2,\ldots,J} \theta_j^*, \tag{5.6}$$

where, for each $j \in \{1, 2, \ldots, J\}$, θ_j^* *is the unique positive solution of the equation*

$$\sum_{i=1}^N \lambda_i \left(e^{R_{ji}\theta} - 1 \right) = \theta.$$

Proof. Consider a fixed $j \in \{1, 2, \ldots, J\}$. Without loss of generality, we assume that $C_j = 1$, by properly normalizing the inequality $(\mathbf{Rx})_j \leq C_j$. In this case, $R_{ji} \leq 1$ for all i, since for each $i \in \{1, 2, \ldots, N\}$, $\mathbf{e}_i \in \mathcal{S} \subset \langle \mathcal{S} \rangle$, and satisfies the constraint $(\mathbf{Re}_i)_j = R_{ji} \leq C_j = 1$.

Now consider the following single-server queueing system. The arrival process is given by the sum $\sum_{i=1}^N R_{ji} A_i(\cdot)$, so arrivals across time slots are independent, and in

each time slot, the amount of work that arrives is $\sum_{i=1}^{N} R_{ji} a_i$, where a_i is a Poisson random variable with mean λ_i, for each i. Note that the arriving amount in a single time slot does not have to be integral. Note also that $\sum_{i=1}^{N} R_{ji}\lambda_i = \tilde{\rho}_j < 1$, since $\rho(\boldsymbol{\lambda}) = \max_j \tilde{\rho}_j < 1$. In each time slot, a unit amount of service is allocated to the total workload in the system. Then, for this system, the workload process $W(\cdot)$ satisfies

$$W(\tau+1) = [W(\tau) - 1]^+ + \sum_{i=1}^{N} R_{ji} a_i(\tau),$$

where $a_i(\tau)$ is the number of arrivals to queue i in the original system in time slot τ. We make two observations for this system. First, $W(\cdot)$ is stochastically dominated by $\sum_{i=1}^{N} R_{ji} Q_i(\cdot)$, where $Q_i(\cdot)$ is the size of queue i in the original system, under any online scheduling policy. This is because for all schedules $\boldsymbol{\sigma} \in \mathcal{S}$, $\boldsymbol{\sigma}$ satisfies $\mathbf{R}\boldsymbol{\sigma} \leq \mathbf{C}$, and hence $\sum_{i=1}^{N} R_{ji}\sigma_i \leq C_j = 1$ for every $\boldsymbol{\sigma} \in \mathcal{S}$. Second, since $R_{ji} \leq 1$ for all i, $\sum_{i=1}^{N} R_{ji} Q_i(\cdot)$ is stochastically dominated by $\sum_{i=1}^{N} Q_i(\cdot)$. Thus, we have

$$\liminf_{L\to\infty} \frac{1}{L} \log \mathbb{P}_{\pi_0}\left(\sum_i Q_i \geq L\right) \geq \liminf_{L\to\infty} \frac{1}{L} \log \mathbb{P}\left(W(\infty) \geq L\right).$$

We now show that

$$\liminf_{L\to\infty} \frac{1}{L} \log \mathbb{P}\left(W(\infty) \geq L\right) \geq -\theta_j^*,$$

where θ_j^* is the unique positive solution of the equation

$$\sum_{i=1}^{N} \lambda_i \left(e^{R_{ji}\theta} - 1\right) = \theta.$$

Consider the log-moment generating function (log-MGF) of the arriving amount in one time slot, given by $\sum_{i=1}^{N} R_{ji} a_i$. Since a_i is a Poisson random variable with mean λ_i for each i, its moment generating function is given by

$$f(\theta) = \exp\left(\sum_{i=1}^{N} \lambda_i (e^{R_{ji}\theta} - 1)\right).$$

Hence, the log-MGF is

$$\log f(\theta) = \sum_{i=1}^{N} \lambda_i (e^{R_{ji}\theta} - 1).$$

By Theorem 1.4 of [23],

$$\lim_{L\to\infty} \frac{1}{L} \log \mathbb{P}\left(W(\infty) \geq L\right) = -\theta_j^*,$$

where $\theta_j^* = \sup\{\theta > 0 : \log f(\theta) < \theta\}$. Since $\log f(\theta) - \theta$ is strictly convex, θ_j^* satisfies

$$\sum_{i=1}^{N} \lambda_i \left(e^{R_{ji}\theta_j^*} - 1\right) = \theta_j^*.$$

$j \in \{1, 2, \ldots, J\}$ is arbitrary, so

$$\liminf_{L\to\infty} \frac{1}{L} \log \mathbb{P}_{\boldsymbol{\pi}_0} \left(\sum_i Q_i \geq L \right) \geq - \min_{j=1,2,\ldots,J} \theta_j^*.$$

□

The following universal lower bound on the tail exponent in input-queued switches is then immediate. As in Section 2.2.1, we use double indexing. Recall that an $n \times n$ input-queued switch has $n^2 = N$ queues.

Corollary 5.3.3 *Consider an $n \times n$ input-queued switch, with an arrival rate vector $\boldsymbol{\lambda}$. Suppose that $\boldsymbol{\lambda}$ is strictly admissible, so that $\rho = \rho(\boldsymbol{\lambda}) < 1$. Consider any policy under which there is a well-defined limiting distribution, and denote this distribution by $\boldsymbol{\pi}_0$. Then, under $\boldsymbol{\pi}_0$,*

$$\liminf_{L\to\infty} \frac{1}{L} \log \mathbb{P}_{\boldsymbol{\pi}_0} \left(\sum_{k,\ell=1}^{n} Q_{k,\ell} \geq L \right) \geq -\theta^*, \tag{5.7}$$

where θ^* is the unique positive solution of the equation $\rho(e^\theta - 1) = \theta$.

Proof. To prove the corollary, consider the representation of the admissible region $\bar{\boldsymbol{\Lambda}}$ in Section 2.2.1. It is of the form $\bar{\boldsymbol{\Lambda}} = \{\mathbf{x} \in [0,1]^{n\times n} : \mathbf{R}\mathbf{x} \leq \mathbf{1}\}$, where all entries of $\mathbf{R}$ are either 0 or 1. In particular, this means that in the notation of Lemma 5.3.2,

for each $j \in \{1, 2, \ldots, J\}$, θ_j^* satisfies

$$\sum_{i=1}^{N} \lambda_i \mathbb{I}_{\{R_{ji}=1\}}(e^{\theta_j^*} - 1) = \widetilde{\rho}_j(e^{\theta_j^*} - 1) = \theta_j^*.$$

Eq. (5.7) is then established by noting $\rho = \max_{j=1,2,\ldots,J} \widetilde{\rho}_j$. □

Eq. (5.7) in Corollary 5.3.3 establishes the optimality of the tail exponent for input-queues switches under **EMUL** (cf. Eq. (5.3) in Theorem 5.3.1). It is also clear from the proof of Corollary 5.3.3 that whenever the incidence matrix $\mathbf{R}$ has entries in $\{0, 1\}$, the tail exponent under **EMUL** is optimal. Input-queued switches are not the only network model that satisfies this condition. For example, the independent-set model of a wireless network also has incidence matrix $\mathbf{R}$ with entries in $\{0, 1\}$. Thus, the tail exponent under **EMUL** is also optimal for the independent-set model of a wireless network.

If $\theta^* > 0$ satisfies the equation $\rho(e^{\theta^*} - 1) = \theta^*$, then $\theta^* \approx \frac{1-\rho}{2\rho}$ when $\rho \approx 1$. This approximation can be derived by considering the first three terms of the Taylor expansion of e^{θ}. If $\theta_j^* > 0$ satisfies the equation

$$\sum_{i=1}^{N} \lambda_i \left(e^{R_{ji}\theta_j^*} - 1\right) = \theta_j^*,$$

then

$$\theta_j^* \leq \frac{2(1 - \tilde{\rho}_j)}{\sum_{i=1}^{N} \lambda_i R_{ji}^2}.$$

(The calculation is elementary and omitted.) Thus, in general, the lower bound on the tail exponent in Lemma 5.3.2 depends on the network structure through $\mathbf{R}$. It is of interest to see whether for general single-hop switched networks, tighter lower bound on the tail exponent can be derived.

Optimality of The Expected Total Queue Size. Here we argue that the scaling of the average total queue size under our policy is optimal for input-queued switches. To that end, as argued in Shah et al. [51], when all input and output ports approach

critical load, the long-run average total queue size under any policy must scale at least as fast as $\sqrt{N}/(1-\rho)$, for any $n \times n$ input-queued switch with $N = n^2$ queues. For completeness, we include the proof of this lower bound here. As in Section 2.2.1, we use double indexing.

Lemma 5.3.4 *Consider an $n \times n$ input-queued switch, with an arrival rate vector $\boldsymbol{\lambda}$. Suppose that the loads on all input and output ports are ρ, i.e., $\sum_{k=1}^{n} \lambda_{k,\ell} = \sum_{m} \lambda_{\ell,m} = \rho$, for all $\ell \in \{1, 2, \ldots, n\}$, where $\rho \in (0, 1)$. Consider any policy under which the queue-size process has a well-defined limiting stationary distribution, and let this distribution be denoted by $\boldsymbol{\pi}_0$. Then under $\boldsymbol{\pi}_0$, we must have*

$$\mathbb{E}_{\boldsymbol{\pi}_0}\left[\sum_{k,\ell=1}^{n} Q_{k,\ell}\right] \geq \frac{n\rho}{2(1-\rho)}.$$

Proof. We consider the sums of queue sizes at each output port, i.e., the quantities $\sum_{k=1}^{n} Q_{k,\ell}$ for each $\ell \in \{1, 2, \ldots, n\}$. Since at most one packet can depart at each time slot, $\sum_{k=1}^{n} Q_{k,\ell}$ stochastically dominates the queue size in an $M/D/1$ system, with arrival rate ρ and deterministic service rate 1. Therefore, for each $\ell \in \{1, 2, \ldots, n\}$,

$$\mathbb{E}_{\boldsymbol{\pi}_0}\left[\sum_{k=1}^{n} Q_{k,\ell}\right] \geq \frac{\rho}{2(1-\rho)}.$$

Here, $\frac{\rho}{2(1-\rho)}$ is the expected queue size in steady state in an $M/D/1$ system. Summing over ℓ gives us the desired bound. □

The optimality in terms of the average total queue size is a direct consequence of Theorem 5.3.1 and Lemma 5.3.4.

Corollary 5.3.5 *Consider the same setup as in Lemma 5.3.4. Then in the heavy-traffic limit $\rho \to 1$, our policy is 2-optimal in terms of the average total queue size. More precisely, consider the expected total queue size in the diffusion scale in steady state, i.e., $(1-\rho)\bar{Q}$. Then,*

$$\limsup_{\rho \to 1} (1-\rho)\bar{Q} \leq n$$

under our policy, and

$$\liminf_{\rho \to 1} (1-\rho)\bar{Q} \geq \frac{n}{2}$$

under any other policy.

Proof. Lemma 5.3.4 implies that

$$\liminf_{\rho \to 1} (1-\rho)\bar{Q} \geq \frac{n}{2}$$

under any policy. For the upper bound, note that by Theorem 5.3.1, under our policy,

$$\bar{Q} \leq \frac{J}{2(1-\rho)} + (N+2)K.$$

For input-queued switches, $J \leq 2n$, as remarked in Section 5.3.2, $N = n^2$, and $K = n$. Therefore, under our policy, the expected total queue size scales as

$$\bar{Q} \leq \frac{n}{1-\rho} + (n^2+2)n. \tag{5.8}$$

Now consider the steady-state heavy-traffic scaling $(1-\rho)\mathbf{Q}$. We have that

$$(1-\rho)\bar{Q} \leq n + (1-\rho)(n^2+2)n. \tag{5.9}$$

The term $(1-\rho)(n^2+2)n$ goes to zero as $\rho \to 1$, and hence under our policy,

$$\limsup_{\rho \to 1} (1-\rho)\bar{Q} \leq n.$$

□

Our policy is not optimal in terms of the average total queue size, in general switched networks. In cases where $J >> N$, the moment bounds for the maximum-weight policy give tighter upper bounds. For more discussion, see Section 5.6.

5.4 Insensitivity in Stochastic Networks

This section recalls the background on insensitive stochastic networks that underlies the main results of this chapter. We shall focus on descriptions of the insensitive bandwidth allocation in so-called bandwidth-sharing networks operating in continuous time. Justifications of claims made in this section are provided in Appendix B.

We consider a bandwidth-sharing network operating in continuous time with capacity constraints. The particular bandwidth-sharing policy of interest is the so-called "store-and-forward allocation (**SFA**)," introduced by Bonald and Proutière [8]. We shall use the **SFA** as an idealized policy to design online scheduling policies for switched networks. We now describe the precise model, the **SFA** policy, and what we know about its performance.

Model. Let time be continuous and indexed by $t \in \mathbb{R}_+$. Consider a network with $J \geq 1$ resources indexed from $1, \ldots, J$. Let there be N routes, and suppose that each *packet* on route i consumes an amount $R_{ji} \geq 0$ of resource j, for each $j \in \{1, 2, \ldots, J\}$. Let $\mathcal{K}$ be the set of all resource-route pairs (j, i) such that route i uses resource j, i.e., $\mathcal{K} = \{(j, i) : R_{ji} > 0\}$. Without loss of generality, we assume that for each $i \in \{1, 2, \ldots, N\}$, $\sum_{j=1}^{J} R_{ji} > 0$. Let $\mathbf{R}$ be the $J \times N$ matrix with entries R_{ji}. Let $\mathbf{C} \in \mathbb{R}_+^J$ be a positive *capacity* vector with components C_j. For each route i, *packets* arrive as an independent Poisson process of rate λ_i. Packets arriving on route i require a unit amount of service, deterministically.

We denote the number of packets on route i at time t by $M_i(t)$, and define the queue-size vector at time t by $\mathbf{M}(t) = [M_i(t)]_{i=1}^N \in \mathbb{Z}_+^N$. Each packet gets service from the network at a rate determined according to a bandwidth-sharing policy. We also denote the total residual workload on route i at time t by $W_i(t)$, and let the vector of residual workload at time t be $\mathbf{W}(t) = [W_i(t)]_{i=1}^N$. Once a packet receives its total (unit) amount of service, it departs the network.

We consider online, myopic bandwidth allocations. That is, the bandwidth allocation at time t only depends on the queue-size vector $\mathbf{M}(t)$. When there are m_i packets

on route i, that is, if the vector of packets is $\mathbf{m} = [m_i]_{i=1}^N$, let the total bandwidth allocated to route i be $\phi_i(\mathbf{m}) \in \mathbb{R}_+$. We consider a processor-sharing policy, so that each packet on route i is served at rate $\phi_i(\mathbf{m})/m_i$, if $m_i > 0$. If $m_i = 0$, let $\phi_i(\mathbf{m}) = 0$. If the *bandwidth vector* $\boldsymbol{\phi}(\mathbf{m}) = [\phi_i(\mathbf{m})]_{i=1}^N$ satisfies the capacity constraints

$$\mathbf{R}\boldsymbol{\phi}(\mathbf{m}) \leq \mathbf{C}, \quad \text{component-wise}, \tag{5.10}$$

for all $\mathbf{m} \in \mathbb{Z}_+^N$ then, in light of Definition 2.2.2, we say that $\boldsymbol{\phi}(\cdot)$ is an *admissible bandwidth allocation*. A Markovian description of the system is given by a process $\mathbf{X}(t)$ which contains the queue-size vector $\mathbf{M}(t)$ along with the residual workloads of the set of packets on each route.

Now, on average, λ_i units of work arrive to route i per unit time. Therefore, in order for the Markov process $\mathbf{X}(\cdot)$ to be positive (Harris) recurrent, it is necessary that

$$\mathbf{R}\boldsymbol{\lambda} \ < \ \mathbf{C}, \quad \text{component-wise}. \tag{5.11}$$

All such $\boldsymbol{\lambda} = [\lambda_i]_{i=1}^N \in \mathbb{R}_+^N$ will be called *strictly admissible*, in the same spirit as the admissible region for a switched network.

Store-and-Forward Allocation (SFA) Policy. We describe the store-and-forward allocation policy that was first considered by Massoulié and later analyzed in the thesis of Proutière [45]. Bonald and Proutière [8] established that it induces product-form stationary distributions and is insensitive with respect to phase-type distributions. This policy is shown to be *insensitive* for general service time distributions, including the deterministic service considered here, by Zachary [66]. The relation between this policy, the proportionally fair allocation, and multiclass queueing networks is discussed in depth by Walton [63] and Kelly et al. [35]. The insensitivity property implies that the invariant measure of the process $\mathbf{M}(t)$ only depends on the parameters $\boldsymbol{\lambda} = [\lambda_i]_{i=1}^N \in \mathbb{R}_+^N$, and the condition that the arrival processes are Poisson, and no other aspects of the stochastic description of the system.

We first give an informal motivation for **SFA**. **SFA** is closely related to quasi-reversible queueing networks. Consider a continuous-time multi-class queueing network (without scheduling constraints) consisting of processor sharing queues indexed by $j \in \{1, ..., J\}$ and job types indexed by the routes $i \in \{1, ..., N\}$. Each route i job has a service requirement R_{ji} at each queue j, and a fixed service capacity C_j is shared between jobs at the queue. Here each job will sequentially visit all the queues (so called store-and-forward) and will visit each queue a fixed number of times. If we assume jobs on each route arrive as a Poisson process, then the resulting queueing network will be stable for all strictly admissible arrival rates. Moreover, in steady state, each queue will be independent, with a queue size that scales, with its load ρ, as $\rho/(1-\rho)$. For further details, see Kelly [34]. So, assuming each queue has equal load, the total number of jobs within the network is of the order $J\rho/(1-\rho)$. In other words, these networks have the stability and queue-size scaling that we require, but do not obey the necessary scheduling constraints (5.10). However, these networks do produce an admissible schedule on average. For this reason, we consider a **SFA** policy which, given the number of jobs on each route, allocates the average rate with which jobs are transferred through this multi-class network. Next, we describe this policy (using notation similar to that used in [35, 63]).

Given $\mathbf{m} \in \mathbb{Z}_+^N$, define

$$U(\mathbf{m}) = \left\{ \widetilde{\mathbf{m}} = (\widetilde{m}_{ji} : (j,i) \in \mathcal{K}) \in \mathbb{Z}_+^{|\mathcal{K}|} : \sum_{j:j\in i} \widetilde{m}_{ji} = m_i \text{ for all } 1 \le i \le N \right\}.$$

Here, by the notation $j \in i$ we mean $R_{ji} > 0$. For each $\widetilde{\mathbf{m}} \in U(\mathbf{m})$, we abuse notation somewhat and define $\widetilde{m}_j = \sum_{i:j\in i} \widetilde{m}_{ji}$, for all $j \le J$. Also define

$$\binom{\widetilde{m}_j}{\widetilde{m}_{ji} : i \ni j} = \frac{\widetilde{m}_j!}{\prod_{i:j\in i}(\widetilde{m}_{ji}!)}.$$

In the above, by $i \ni j$ we mean that $R_{ji} > 0$; the notation $i \ni j$ is used when we consider a collection of i satisfying this condition for a given j. For $\mathbf{m} \in \mathbb{Z}_+^N$, we

define $\Phi(\mathbf{m})$ as

$$\Phi(\mathbf{m}) = \sum_{\tilde{\mathbf{m}} \in U(\mathbf{m})} \prod_{j \in J} \left(\binom{\tilde{m}_j}{\tilde{m}_{ji} : i \ni j} \prod_{i : j \in i} \left(\frac{R_{ji}}{C_j} \right)^{\tilde{m}_{ji}} \right). \tag{5.12}$$

We shall define $\Phi(\mathbf{m}) = 0$ if any of the components of $\mathbf{m}$ is negative. The store-and-forward allocation (**SFA**) assigns rates according to the function $\boldsymbol{\phi} : \mathbb{Z}_+^N \to \mathbb{R}_+^N$ so that for any $\mathbf{m} \in \mathbb{Z}_+^N$, $\boldsymbol{\phi}(\mathbf{m}) = (\phi_i(\mathbf{m}))_{i=1}^N$, with

$$\phi_i(\mathbf{m}) = \frac{\Phi(\mathbf{m} - \mathbf{e}_i)}{\Phi(\mathbf{m})}, \tag{5.13}$$

where, we recall that $\mathbf{m} - \mathbf{e}_i$ is the same as $\mathbf{m}$ at all but the ith component; its ith component equals $m_i - 1$. The bandwidth allocation $\boldsymbol{\phi}(\mathbf{m})$ is the stationary throughput of jobs on the routes of a multi-class queueing network (described above), conditional on there being $\mathbf{m}$ jobs on each route.

A priori it is not clear if the above described bandwidth allocation is even admissible (i.e., satisfies (5.10)). This can be argued as follows. The $\boldsymbol{\phi}(\mathbf{m})$ can be related to the stationary throughput of a closed multi-class network with a finite number of jobs, $\mathbf{m}$, on each route. Under this scenario (due to finite number of jobs), each queue must be stable. Therefore, the load on each queue, $\mathbf{R}\boldsymbol{\phi}(\mathbf{m})$, must be less than the overall system capacity $\mathbf{C}$. That is, the allocation is admissible. The precise argument along these lines is provided in, for example [35, Corollary 2] and [63, Lemma 4.1].

The **SFA** policy induces a product-form invariant distribution for the number of packets waiting in the bandwidth-sharing network and is insensitive. We summarize this in the following result.

Theorem 5.4.1 *Consider a bandwidth-sharing network with* $\mathbf{R}\boldsymbol{\lambda} < \mathbf{C}$. *Under the* **SFA** *policy described above, the Markov process* $\mathbf{X}(t)$ *is positive (Harris) recurrent and* $\mathbf{M}(t)$ *has a unique stationary probability distribution* $\boldsymbol{\pi}$ *given by*

$$\boldsymbol{\pi}(\mathbf{m}) = \frac{\Phi(\mathbf{m})}{\Phi} \prod_{i=1}^{N} \lambda_i^{m_i}, \quad \textit{for all } \mathbf{m} \in \mathbb{Z}_+^N, \tag{5.14}$$

where

$$\Phi = \prod_{j=1}^{J} \left(\frac{C_j}{C_j - \sum_{i:i \ni j} R_{ji} \lambda_i} \right) \tag{5.15}$$

is a normalizing factor. Furthermore, the steady-state residual workload of packets waiting in the network can be characterized as follows. First, the steady-state distribution of the residual workload of a packet is independent from $\boldsymbol{\pi}$*. Second, in steady state, conditioned on the number of packets on each route of the network, the residual workload of each packet is uniformly distributed on* $[0, 1]$*, and is independent from the residual workloads of other packets.*

Statements similar to Theorem 5.4.1 have appeared in previous works; for example, [7], [63, Proposition 4.2] and [35]. Theorem 5.4.1 is a summary of these statements, and, for completeness, it is proved in Appendix B.

The following property of the stationary distribution $\boldsymbol{\pi}$ described in Theorem 5.4.1 that will be useful.

Proposition 5.4.2 *Consider the setup of Theorem 5.4.1 and let* $\boldsymbol{\pi}$ *be as described by* (5.14)*. Define a measure* $\widetilde{\boldsymbol{\pi}}$ *on* $\mathbb{Z}_+^{|\mathcal{K}|}$ *as follows: for* $\widetilde{\mathbf{m}} \in \mathbb{Z}_+^{|\mathcal{K}|}$*,*

$$\widetilde{\boldsymbol{\pi}}(\widetilde{\mathbf{m}}) = \frac{1}{\Phi} \prod_{j=1}^{J} \left(\binom{\tilde{m}_j}{\tilde{m}_{ji} : i \ni j} \prod_{i:j \in i} \left(\frac{R_{ji} \lambda_i}{C_j} \right)^{\tilde{m}_{ji}} \right). \tag{5.16}$$

Then, for any $L \in \mathbb{Z}_+$*,*

$$\boldsymbol{\pi}\left(\left\{ \mathbf{m} : \sum_{i=1}^{N} m_i = L \right\} \right) = \widetilde{\boldsymbol{\pi}}\left(\left\{ \widetilde{\mathbf{m}} : \sum_{j=1}^{J} \tilde{m}_j = L \right\} \right). \tag{5.17}$$

We relate the distribution $\widetilde{\boldsymbol{\pi}}$ to the stationary distribution of an insensitive multi-class queueing network with a product-form stationary distribution and geometrically distributed queue sizes.

Proposition 5.4.3 *Consider the distribution* $\widetilde{\boldsymbol{\pi}}$ *defined in* (5.16)*. Then, for any*

$\mathbf{L} = (L_1, \ldots, L_J) \in \mathbb{Z}_+^J$,

$$\widetilde{\boldsymbol{\pi}}\big(\widetilde{m}_1 = L_1, \ldots, \widetilde{m}_J = L_J\big) \overset{(a)}{=} \sum_{(\tilde{m}_{ji}) \in U(\mathbf{L})} \widetilde{\boldsymbol{\pi}}((\tilde{m}_{ji})) = \prod_{j=1}^{J} \widetilde{\rho}_j^{L_j}(1 - \widetilde{\rho}_j), \tag{5.18}$$

where $\widetilde{\rho}_j = \big(\sum_{i:i\ni j} R_{ji}\lambda_j\big)/C_j$.

Using Theorem 5.4.1, Propositions 5.4.2 and 5.4.3, we can compute the expected value and the probability tail exponent of the steady-state total residual workload in the system. Recall that the total residual workload in the system at time t is given by $\sum_{i=1}^{N} W_i(t)$.

Proposition 5.4.4 *Consider a bandwidth-sharing network with* $\mathbf{R\lambda} < \mathbf{C}$, *operating under the* **SFA** *policy. Denote the load induced by* $\boldsymbol{\lambda}$ *by* $\rho = \rho(\boldsymbol{\lambda})(< 1)$, *and for each* j, *let* $\widetilde{\rho}_j = \left(\sum_i R_{ji}\lambda_i\right)/C_j$. *Then* $\mathbf{W}(\cdot)$ *has a unique stationary probability distribution. With respect to this stationary distribution, the following properties hold.*

(i) The expected total residual workload is given by

$$\mathbb{E}\left[\sum_{i=1}^{N} W_i\right] = \frac{1}{2}\sum_{j=1}^{J} \frac{\widetilde{\rho}_j}{1 - \widetilde{\rho}_j}. \tag{5.19}$$

(ii) The distribution of the total residual workload has an exponential tail with exponent given by

$$\lim_{L\to\infty} \frac{1}{L} \log \mathbb{P}\left(\sum_{i=1}^{N} W_i \geq L\right) = -\theta^*, \tag{5.20}$$

where θ^* *is the unique* positive *solution of the equation* $\rho(e^{\theta} - 1) = \theta$.

5.5 The Policy and Its Performance

5.5.1 EMUL for switched networks

Given a switched network, denoted by **SN**, with schedule set $\mathcal{S}$ and N queues, let $\langle \mathcal{S} \rangle$ have rank J and representation (cf. (5.1))

$$\langle \mathcal{S} \rangle = \{ \mathbf{x} \in [0,1]^N : \mathbf{R}\mathbf{x} \leq \mathbf{C} \}, \quad \mathbf{R} \in \mathbb{R}_+^{J \times N}, \ \mathbf{C} \in \mathbb{R}_+^J.$$

As mentioned in Section 5.3, we couple **SN** with an appropriate bandwidth-sharing network. Consider the following virtual bandwidth-sharing network, denoted by **BN**, with N routes corresponding to each of these N queues. The resource-route relation is determined by the same matrix $\mathbf{R}$; and the J resources have capacities given by $\mathbf{C}$. The networks **SN** and **BN** are coupled by having identical arrivals. That is, a packet arrives to queue i in **SN** iff a packet arrives to route i in **BN** at the same time.

Now **EMUL** for **SN** will be derived from **BN**. Specifically, let **BN** operate under the insensitive **SFA** policy described in Section 5.4. By Theorem 5.4.1 and Proposition 5.4.2, **SFA** induces a desirable stationary distribution of queue sizes in **BN**. If we could use the rate allocation **SFA** directly in **SN**, it would give us desired performance bounds on the stationary queue sizes, in **SN**. However, the instantaneous rate allocations under **SFA** change all the time, and are only required to utilize points inside $\langle \mathcal{S} \rangle (= \bar{\mathbf{\Lambda}})$, not necessarily points in $\mathcal{S}$. In contrast, in **SN** the rate allocation can change only once per discrete time slot and it must always employ a schedule from $\mathcal{S}$. The key to **EMUL**, then, is an effective way to emulate in an online manner the rate allocation of **BN** under **SFA**, taking into account the scheduling constraints $\mathcal{S}$ and the discrete-time constraint.

To that end, we describe this emulation policy. We start by introducing some useful notation. Let $\mathbf{A}(\cdot) = \big(A_i(\cdot)\big)$ be the vector of exogenous, independent Poisson processes according to which unit-sized packets arrive to both **BN** and **SN**, simultaneously. Recall that $A_i(\cdot)$ is a Poisson process with rate λ_i. Let $\mathbf{M}(t) = \big(M_i(t)\big)_{i=1}^N$ denote the vector of numbers of packets waiting on the N routes in **BN** at time

$t \geq 0$. In **BN**, the services are allocated according to the **SFA** policy described in Section 5.4. Let $\Lambda_i^{\text{SFA}}(t)$ denote the total amount of service allocated to all packets on route i during the interval $[0,t]$, for $t \geq 0$, with $\Lambda_i^{\text{SFA}}(0) = 0$ for $1 \leq i \leq N$, and let $\mathbf{\Lambda}^{\text{SFA}}(\cdot) = \big(\Lambda_i^{\text{SFA}}(\cdot)\big)_{i=1}^N$. By definition, all components of $\mathbf{\Lambda}^{\text{SFA}}(\cdot)$ are non-decreasing and Lipschitz continuous. Furthermore, $(\mathbf{\Lambda}^{\text{SFA}}(t+s) - \mathbf{\Lambda}^{\text{SFA}}(t))/s \in \langle \mathcal{S} \rangle$ for any $t \geq 0$ and $s > 0$. Recall that the (right-)derivative of $\mathbf{\Lambda}^{\text{SFA}}(\cdot)$ is determined by $\mathbf{M}(\cdot)$ through the function $\boldsymbol{\phi}(\cdot)$ as defined in (5.13).

Now we describe the policy for **SN**, which will rely on $\mathbf{\Lambda}^{\text{SFA}}(\cdot)$. Let $\mathbf{S}(\tau) = \big(S_i(\tau)\big)$ denote the cumulative amount of service allocated in **SN** by the policy up to the end of time slot $\tau - 1$, with $\mathbf{S}(0) = \mathbf{0}$. **EMUL** determines how $\mathbf{S}(\cdot)$ is updated. Let $\mathbf{Q}(\tau) = \big(Q_i(\tau)\big)$ be the queue sizes measured at the end of time slot τ. Then, **EMUL** decides the schedule $\boldsymbol{\sigma}(\tau) = \mathbf{S}(\tau+1) - \mathbf{S}(\tau) \in \mathcal{S}$ at the very end of time slot τ, right after the queue-size information $\mathbf{Q}(\tau)$ is updated. This decision is made as follows. Let $\mathbf{D}(\tau) = \mathbf{\Lambda}^{\text{SFA}}(\tau) - \mathbf{S}(\tau)$. Let $\rho(\mathbf{D}(\tau))$ be the optimal objective value in the optimization problem PRIMAL($\mathbf{D}(\tau)$) defined in (2.14). In particular, there exists a non-negative combination of schedules in $\mathcal{S}$ such that

$$\sum_{\boldsymbol{\sigma} \in \mathcal{S}} \tilde{\alpha}_{\boldsymbol{\sigma}} \boldsymbol{\sigma} \geq \mathbf{D}(\tau), \quad \text{and} \quad \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \tilde{\alpha}_{\boldsymbol{\sigma}} = \rho(\mathbf{D}(\tau)). \tag{5.21}$$

We claim that in fact, we can find non-negative numbers $\alpha_{\boldsymbol{\sigma}}$, $\boldsymbol{\sigma} \in \mathcal{S}$, such that

$$\sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} \boldsymbol{\sigma} = \mathbf{D}(\tau), \quad \text{and} \quad \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} = \rho(\mathbf{D}(\tau)). \tag{5.22}$$

This is formalized in the following lemma.

Lemma 5.5.1 *Let $\mathbf{D} \in \mathbb{R}_+^N$ be a non-negative vector. Consider the static planning problem* PRIMAL($\mathbf{D}$) *defined in* (2.14). *Let the optimal objective value to* PRIMAL($\mathbf{D}$) *be $\rho(\mathbf{D})$. Then there exist $\alpha_{\boldsymbol{\sigma}} \geq 0$, $\boldsymbol{\sigma} \in \mathcal{S}$, such that* (5.22) *hold.*

The proof of the lemma relies on Assumption 5.2.1, and is provided in Appendix B.

There could be many possible non-negative combinations of $\mathbf{D}(\tau)$ satisfying (5.22). If there exist non-negative numbers $\alpha_{\boldsymbol{\sigma}}$, $\boldsymbol{\sigma} \in \mathcal{S}$, satisfying (5.22) with $\alpha_{\boldsymbol{\sigma}'} \geq 1$ for

some $\boldsymbol{\sigma}' \in \mathcal{S}$, then choose $\boldsymbol{\sigma}'$ as the schedule: set $\boldsymbol{\sigma}(\tau) = \boldsymbol{\sigma}'$. If no such decomposition exists for $\mathbf{D}(\tau)$, then set $\boldsymbol{\sigma}(\tau) = \widetilde{\boldsymbol{\sigma}}$, where $\widetilde{\boldsymbol{\sigma}}$ is a solution (ties broken arbitrarily) of

$$\text{maximize} \quad \sum_i \sigma_i \quad \text{over} \quad \boldsymbol{\sigma} \in \mathcal{S},\ \boldsymbol{\sigma} \leq \mathbf{D}(\tau). \tag{5.23}$$

Note that $\mathbf{0}$ is a feasible solution for the above problem as $\mathbf{0} \in \mathcal{S}$ and $\mathbf{0} \leq \mathbf{D}(\tau)$. Note also that for all times τ, $\boldsymbol{\sigma}(\tau) \leq \mathbf{D}(\tau)$.

The above is a complete description of **EMUL**. Observe that it is an online policy, as the virtual network **BN** can be simulated in an online manner, and, given this, the allocation decisions in **SN** rely only on the history of **BN** and **SN**.

5.5.2 Proof of the Main Theorem (Theorem 5.3.1)

The proof is divided in three parts. The first part describes a sample-path-wise relation between $\mathbf{Q}(\cdot)$ and $\mathbf{M}(\cdot)$, which implies that $\mathbf{Q}(\cdot)$ is essentially dominated by $\mathbf{M}(\cdot)$ at all times. Note that this domination is a distribution-free statement. The second part utilizes this fact to establish the positive recurrence of the **SN** Markov chain. The third part is a consequence of the first two parts, and using Theorem 5.4.1, establishes the quantitative claims in Theorem 5.3.1.

Part 1. Dominance. We start by establishing that the queue sizes $\mathbf{Q}(\cdot)$ of **SN** are effectively dominated by the workloads $\mathbf{W}(\cdot)$ of **BN** at all times. We state this result formally in Proposition 5.5.4, which is a consequence of Lemmas 5.5.2 and 5.5.3 below.

Lemma 5.5.2 *Consider the evolution of queue sizes in the **BN** and **SN** networks, fed by identical arrival process. Initially, $\mathbf{Q}(0) = \mathbf{M}(0) = \mathbf{0}$. Let $\mathbf{W}(\tau) = (W_i(\tau))$ denote the amount of unfinished work in all N queues under the **BN** network at time τ. Then for any $\tau \geq 0$ and $1 \leq i \leq N$,*

$$W_i(\tau) \leq Q_i(\tau) \leq W_i(\tau) + D_i(\tau) \leq M_i(\tau) + D_i(\tau), \tag{5.24}$$

where $\mathbf{D}(\tau) = \Lambda^{SFA}(\tau) - \mathbf{S}(\tau)$ is as described in Section 5.5.1.

Proof. Consider any $i \in \{1, 2, \ldots, N\}$ and $\tau \geq 0$. From (2.2), in **SN**,

$$Q_i(\tau) = A_i(\tau) - S_i(\tau) + Z_i(\tau), \tag{5.25}$$

where $Z_i(\tau)$ is the cumulative amount of idling at the ith queue in **SN**. In a similar manner, in **BN**,

$$W_i(\tau) = A_i(\tau) - \Lambda_i^{\text{SFA}}(\tau) + \widehat{Z}_i(\tau), \tag{5.26}$$

where $\widehat{Z}_i(\tau)$ is the cumulative amount of idling for the ith queue in **BN**. Since by construction, $\mathbf{D}(\tau) = \Lambda^{\text{SFA}}(\tau) - \mathbf{S}(\tau)$, and $\mathbf{D}(\tau) \geq \mathbf{0}$, we have that

$$S_i(\tau) \leq \Lambda_i^{\text{SFA}}(\tau) \ \leq S_i(\tau) + D_i(\tau). \tag{5.27}$$

By definition, the instantaneous rate allocation to the ith queue satisfies $\frac{d}{dt^+}\Lambda_i^{\text{SFA}}(t) = 0$ if $W_i(t) = 0$ (equivalently, if $M_i(t) = 0$) for any $t \geq 0$. Therefore, for all i and τ, $\widehat{Z}_i(\tau) = 0$, and $W_i(\tau) = A_i(\tau) - \Lambda_i^{\text{SFA}}(\tau)$. On the other hand, by Skorohod's map,

$$\begin{aligned} Z_i(\tau) &= \sup_{0 \leq s \leq \tau} \left[S_i(s) - A_i(s)\right]^+ \\ &\leq \sup_{0 \leq s \leq \tau} \left[\Lambda_i^{\text{SFA}}(s) - A_i(s)\right]^+ \\ &= \widehat{Z}_i(\tau) \ = \ 0, \end{aligned} \tag{5.28}$$

hence for all i and τ, $Z_i(\tau) = 0$, and $Q_i(\tau) = A_i(\tau) - S_i(\tau)$. It then follows that

$$\begin{aligned} Q_i(\tau) &= A_i(\tau) - S_i(\tau) \\ &\leq A_i(\tau) - \Lambda_i^{\text{SFA}}(\tau) + D_i(\tau) \\ &= W_i(\tau) + D_i(\tau), \end{aligned} \tag{5.29}$$

and

$$W_i(\tau) = A_i(\tau) - \Lambda_i^{\text{SFA}}(\tau) \leq A_i(\tau) - S_i(\tau) = Q_i(\tau).$$

Since the workload at the ith queue equals the total amount of unfinished work for all of the $M_i(\tau)$ packets waiting at the ith queue, and since each packet has at most a unit amout of unfinished work, $W_i(\tau) \leq M_i(\tau)$. □

Lemma 5.5.3 *Let* $\mathbf{D}(\tau)$ *be as in Lemma 5.5.2. For all* $\tau \geq 0$, $\rho(\mathbf{D}(\tau)) \leq N+2$. *In particular,*

$$\sum_i D_i(\tau) \leq K(N+2), \quad \textit{where} \quad K = \max_{\boldsymbol{\sigma} \in \mathcal{S}} \sum_i \sigma_i. \tag{5.30}$$

Proof. This result is established as follows. First, observe that $\mathbf{D}(0) = \mathbf{0}$ and therefore $\rho(\mathbf{D}(0)) = 0$. Next, we show that $\rho(\mathbf{D}(\tau+1)) \leq \rho(\mathbf{D}(\tau)) + 1$. That is, $\rho(\mathbf{D}(\cdot))$ can at most increase by 1 in each time slot. And finally, we show that it cannot increase once it exceeds $N+1$. That is, if $\rho(\mathbf{D}(\tau)) \geq N+1$, then $\rho(\mathbf{D}(\tau+1)) \leq \rho(\mathbf{D}(\tau))$. This will complete the proof.

We start by establishing that $\rho(\mathbf{D}(\cdot))$ increases by at most 1 in unit time. By definition,

$$\begin{aligned}
\mathbf{D}(\tau+1) &= \Lambda^{\text{SFA}}(\tau+1) - \mathbf{S}(\tau+1) \\
&= \Lambda^{\text{SFA}}(\tau) - \mathbf{S}(\tau) + \Big(\Lambda^{\text{SFA}}(\tau+1) - \Lambda^{\text{SFA}}(\tau) - \boldsymbol{\sigma}(\tau)\Big) \\
&= \mathbf{D}(\tau) + d\Lambda^{\text{SFA}}(\tau) - \boldsymbol{\sigma}(\tau) \\
&= \Big(\mathbf{D}(\tau) - \boldsymbol{\sigma}(\tau)\Big) + d\Lambda^{\text{SFA}}(\tau),
\end{aligned} \tag{5.31}$$

where $d\Lambda^{\text{SFA}}(\tau) = \Lambda^{\text{SFA}}(\tau+1) - \Lambda^{\text{SFA}}(\tau)$. As remarked earlier, $\boldsymbol{\sigma}(\tau) \leq \mathbf{D}(\tau)$ component-wise. Therefore, by (2.9) it follows that

$$\rho\big(\mathbf{D}(\tau+1)\big) \leq \rho\big(\mathbf{D}(\tau) - \boldsymbol{\sigma}(\tau)\big) + \rho(d\Lambda^{\text{SFA}}(\tau)).$$

Note that $\rho(d\Lambda^{\text{SFA}}(\tau)) \leq 1$ because the instantaneous service rate under **SFA** is always admissible. Since $\mathbf{D}(\tau) \geq \mathbf{D}(\tau) - \boldsymbol{\sigma}(\tau) \geq \mathbf{0}$, any feasible solution to PRIMAL $\big(\mathbf{D}(\tau)\big)$

is also feasible to PRIMAL $\big(\mathbf{D}(\tau) - \boldsymbol{\sigma}(\tau)\big)$, and hence

$$\rho\big(\mathbf{D}(\tau) - \boldsymbol{\sigma}(\tau)\big) \leq \rho\big(\mathbf{D}(\tau)\big).$$

Hence,

$$\rho\big(\mathbf{D}(\tau+1)\big) \leq \rho(\mathbf{D}(\tau)) + 1. \tag{5.32}$$

Next, we shall argue that if $\rho(\mathbf{D}(\tau)) \geq N+1$, then $\rho(\mathbf{D}(\tau+1)) \leq \rho(\mathbf{D}(\tau))$. To that end, suppose that $\rho(\mathbf{D}(\tau)) \geq N+1$. Now $\frac{1}{\rho(\mathbf{D}(\tau))}\mathbf{D}(\tau) \in \langle \mathcal{S} \rangle$. Note that $\langle \mathcal{S} \rangle$ is a convex set in a N-dimensional space with extreme points contained in $\mathcal{S}$. Therefore, by Carathéodory's theorem, $\frac{1}{\rho(\mathbf{D}(\tau))}\mathbf{D}(\tau)$ can be written as a convex combination of at most $N+1$ elements in $\mathcal{S}$. That is, there exist $\alpha_k \geq 0$ with $\sum_{k=1}^{N+1} \alpha_k = 1$, and $\boldsymbol{\sigma}^k \in \mathcal{S}$, $k \in \{1, 2, \ldots, N+1\}$, such that

$$\frac{1}{\rho(\mathbf{D}(\tau))}\mathbf{D}(\tau) = \sum_{k=1}^{N+1} \alpha_k \boldsymbol{\sigma}^k. \tag{5.33}$$

It follows that there exists some $k^* \in \{1, 2, \ldots, N+1\}$, such that $\alpha_{k^*} \geq 1/(N+1)$. Since $\rho(\mathbf{D}(\tau)) \geq N+1$, $\rho(\mathbf{D}(\tau))\alpha_{k^*} \geq 1$. That is, $\mathbf{D}(\tau)$ can be written as a convex combination of elements from $\mathcal{S}$ with one of them, $\boldsymbol{\sigma}^{k^*}$, having an associated coefficient that satisfies $\rho(\mathbf{D}(\tau))\alpha_{k^*} \geq 1$, as required. In this case, we have

$$\mathbf{D}(\tau) - \boldsymbol{\sigma}^{k^*} = \sum_{k=1, k\neq k^*}^{N+1} \rho(\mathbf{D}(\tau))\alpha_k \boldsymbol{\sigma}^k + (\rho(\mathbf{D}(\tau))\alpha_{k^*} - 1)\boldsymbol{\sigma}^{k^*}. \tag{5.34}$$

Therefore,

$$\rho\Big(\mathbf{D}(\tau) - \boldsymbol{\sigma}^{k^*}\Big) \leq \rho(\mathbf{D}(\tau)) - 1. \tag{5.35}$$

Our scheduling policy chooses such a schedule; that is, $\boldsymbol{\sigma}(\tau) = \boldsymbol{\sigma}^{k^*}$. Therefore,

$$\mathbf{D}(\tau+1) = \mathbf{D}(\tau) - \boldsymbol{\sigma}^{k^*} + d\Lambda^{\mathrm{SFA}}(\tau). \tag{5.36}$$

By another application of (2.9) it follows that

$$\begin{aligned}\rho(\mathbf{D}(\tau+1)) &\leq \rho\big(\mathbf{D}(\tau)-\boldsymbol{\sigma}^{k^*}\big)+\rho(d\Lambda^{\mathrm{SFA}}(\tau))\\ &\leq \rho\big(\mathbf{D}(\tau)\big)-1+1,\\ &= \rho\big(\mathbf{D}(\tau)\big),\end{aligned} \tag{5.37}$$

where again we have used the fact that $\rho(d\Lambda^{\mathrm{SFA}}(\tau)) \leq 1$, due to the feasibility of **SFA** policy and (5.35). This establishes that $\rho(\mathbf{D}(\tau)) \leq N+2$ for all $\tau \geq 0$. That is, for each $\tau \geq 0$, there exist $\alpha_{\boldsymbol{\sigma}} \geq 0$ for all $\boldsymbol{\sigma} \in \mathcal{S}$, such that $\sum_{\boldsymbol{\sigma}} \rho(\mathbf{D}(\tau))\alpha_{\boldsymbol{\sigma}} \leq N+2$ and

$$\mathbf{D}(\tau) \leq \sum_{\boldsymbol{\sigma}} \alpha_{\boldsymbol{\sigma}}\boldsymbol{\sigma}. \tag{5.38}$$

Therefore,

$$\begin{aligned}\sum_i D_i(\tau) &= \mathbf{D}(\tau)\cdot\mathbf{1}\\ &\leq \sum_{\boldsymbol{\sigma}} \rho(\mathbf{D}(\tau))\alpha_{\boldsymbol{\sigma}}\boldsymbol{\sigma}\cdot\mathbf{1}\\ &\leq \Big(\sum_{\boldsymbol{\sigma}} \rho(\mathbf{D}(\tau))\alpha_{\boldsymbol{\sigma}}\Big)\Big(\max_{\boldsymbol{\sigma}\in\mathcal{S}}\sum_i \sigma_i\Big)\\ &\leq (N+2)K,\end{aligned} \tag{5.39}$$

where $K = \max_{\boldsymbol{\sigma}\in\mathcal{S}}\sum_i \sigma_i$. This completes the proof of Lemma 5.5.3. □

Lemmas 5.5.2 and 5.5.3 together imply the following proposition.

Proposition 5.5.4 *Let* $\mathbf{Q}(\cdot)$ *and* $\mathbf{W}(\cdot)$ *be as in Lemma 5.5.2. Then for all time* τ*,*

$$\sum_{i=1}^N W_i(\tau) \leq \sum_{i=1}^N Q_i(\tau) \leq \sum_{i=1}^N W_i(\tau) + K(N+2) \leq \sum_{i=1}^N M_i(\tau) + K(N+2), \tag{5.40}$$

where $K = \max_{\boldsymbol{\sigma}\in\mathcal{S}}\Big(\sum_{i=1}^N \sigma_i\Big)$.

Proof. We obtain the bounds (5.40) by summing inequality (5.24) over $i \in \{1,2,\ldots,N\}$, and using the bound (5.30). □

Part 2. Positive recurrence. We start by defining the Markov chain describing the system evolution under the policy of interest. There are essentially two systems that evolve in a coupled manner under our policy: the virtual bandwidth-sharing network **BN** and the switched network **SN** of interest. The two networks are fed by the same arrival processes which are exogenous and Poisson (and hence Markov). The virtual system **BN** has a Markovian state consisting of the packets whose services are not completed, represented by the vector $\mathbf{M}(\cdot)$, and their residual services. The residual services of $M_i(\cdot)$ packets queued on route i can be represented by a non-negative, finite measure $\mu_i(\cdot)$ on $[0,1]$: unit mass is placed at each of the points $0 \leq s_1, \ldots, s_{M_i(t)} \leq 1$ if the unfinished work of $M_i(t)$ packets are given by $0 < s_1, \ldots, s_{M_i(t)} \leq 1$.

We now consider a Markovian description of the network **SN** in discrete time: let $\mathbf{X}(\tau)$ be the state of the system defined as

$$\mathbf{X}(\tau) = (\mathbf{M}(\tau), \boldsymbol{\mu}(\tau), \mathbf{Q}(\tau), \mathbf{D}(\tau)), \tag{5.41}$$

where $(\mathbf{M}(\tau), \boldsymbol{\mu}(\tau))$ represents the state of **BN** at time τ, $\mathbf{Q}(\tau)$ is the vector of queue sizes in **SN** at time τ and $\mathbf{D}(\tau)$ is the "difference" vector maintained by the scheduling policy for **SN**, as described in Section 5.5.1. Clearly, $\mathbf{X}(\cdot)$ is Markov. We now define the state space X of the Markov chain $\mathbf{X}(\cdot)$. Informally speaking, X will consist of points that not only can be reached from the zero state $\mathbf{0}$, but also can reach $\mathbf{0}$, in finite time. More precisely, first note that X is a subset of the product space

$$\mathbb{Z}_+^N \times \mathcal{M}([0,1])^N \times \mathbb{Z}_+^N \times \mathbb{R}_+^N,$$

where $\mathcal{M}([0,1])$ is the space of all non-negative, finite measures on $[0,1]$. We endow $\mathcal{M}([0,1])$ with the weak topology, which is induced by the Prohorov's metric. This results in a complete and separable metric (Polish) space. The other spaces $\mathbb{Z}_+$ and $\mathbb{R}_+$ are endowed by obvious metrics (e.g., ℓ_1). The entire product space is endowed with metric that is maximum of the metrics on the component spaces. The resulting product space is Polish, on which a Borel σ-algebra can be defined. Then, X consists

of points $\mathbf{x}$ in this product space such that

$$\mathbb{P}_{\mathbf{x}}(T_{\mathbf{0}} < \infty) = 1, \quad \text{and} \quad \mathbb{P}_{\mathbf{0}}(T_{\mathbf{x}} < \infty) = 1.$$

Here $\mathbb{P}_{\mathbf{x}}(A) \equiv \mathbb{P}(A \mid \mathbf{X}(0) = \mathbf{x})$ denotes probability conditioned on the initial condition $\mathbf{X}(0) = \mathbf{x}$, and for a measurable set A, and $T_A = \inf\{\tau \geq 1 : \mathbf{X}(\tau) \in A\}$ denotes the return time to A.

Given the Markovian description $\mathbf{X}(\tau)$ of **SN**, we establish its positive (Harris) recurrence in the following lemma.

Lemma 5.5.5 *Consider a switched network **SN** with a strictly admissible arrival rate vector $\boldsymbol{\lambda}$, with $\rho(\boldsymbol{\lambda}) < 1$. Suppose that at time 0, the system is empty. Let $\mathbf{X}(\cdot)$ be as defined in Eq.* (5.41)*. Then $\mathbf{X}(\cdot)$ is positive recurrent.*

The proof of the lemma is technical, and is deferred to Appendix B. The idea is that the evolution of **BN** is not affected by **SN**, and that **BN** is, on its own, positive recurrent. Hence, starting from any initial state, the Markov process $(\mathbf{M}(\cdot), \boldsymbol{\mu}(\cdot))$ that describes the evolution of **BN**, reaches the null state, i.e., $(\mathbf{M}(\cdot), \boldsymbol{\mu}(\cdot)) = \mathbf{0}$ at some finite expected time. Once **BN** reaches the null state, it stays at this state for an arbitrarily large amount of time with positive probability. By our policy, $\mathbf{Q}(\cdot)$ and $\mathbf{D}(\cdot)$ can be driven to $\mathbf{0}$ within this time interval. This establishes that $\mathbf{X}(\cdot)$ reaches the null state in finite expected time, and that $\mathbf{X}(\cdot)$ is positive recurrent.

Part 3. Completing the proof. The positive recurrence of the Markov chain $\mathbf{X}(\cdot)$ implies that it possesses a unique stationary distribution and that it is ergodic. Let $\bar{W} = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{i=1}^{N} W_i\right]$, where, similar to Lemma 5.5.2, W_i is the steady-state workload on queue i in **BN**. Define $\bar{M}$ similarly. By ergodicity, the time average of the total queue size equals the expected total queue size in steady state, i.e., $\bar{Q}$, and similarly for $\bar{W}$. Therefore, by Proposition 5.5.4,

$$\bar{Q} \leq \bar{W} + K(N+2).$$

By Proposition 5.4.4,

$$\bar{W} = \frac{1}{2}\left(\sum_{j=1}^{J} \frac{\tilde{\rho}_j}{1-\tilde{\rho}_j}\right).$$

Thus,

$$\bar{Q} \leq \bar{W} + K(N+2) = \frac{1}{2}\left(\sum_{j=1}^{J} \frac{\tilde{\rho}_j}{1-\tilde{\rho}_j}\right) + K(N+2).$$

We now establish the tail exponent in (5.3). By Proposition 5.5.4,

$$\sum_{i=1}^{N} W_i(\tau) \leq \sum_{i=1}^{N} Q_i(\tau) \leq \sum_{i=1}^{N} W_i(\tau) + K(N+2),$$

deterministically and for all times τ. Since $K(N+2)$ is a constant, $\sum_{i=1}^{N} Q_i(\cdot)$ and $\sum_{i=1}^{N} W_i(\cdot)$ have the same tail exponent in steady state. By Proposition 5.4.4, the tail exponent $\beta(\mathbf{W})$ of $\sum_{i=1}^{N} W_i$ in steady state is given by

$$\beta(\mathbf{W}) = -\theta^*,$$

where θ^* is the unique positive solution of the equation $\rho(e^{\theta} - 1) = \theta$, so

$$\beta(\mathbf{Q}) = \beta(\mathbf{W}) = -\theta^*.$$

5.6 Discussion

We presented a novel scheduling policy **EMUL** for a generic single-hop switched network model. The policy, in effect, emulates the so-called Store-and-forward (**SFA**) continuous-time bandwidth-sharing policy. The insensitivity property of **SFA** along with the relation of its stationary distribution with that of multi-class queueing networks leads to the explicit characterization of the stationary distribution of queue sizes induced by our policy. This allows us to establish the optimality of our policy in

terms of the tail exponent and that with respect to the average total queue size for a class of switched networks, including input-queued switches. As a consequence, this settles a conjecture stated in [51]. On the technical end, a key contribution in the chapter is designing a discrete-time scheduling policy by emulating a continuous-time rate allocation policy, and this may be of independent interest in other domains of applications. We also remark that the idea of designing a discrete-time policy by emulating a continuous-time policy is not new; for example, such emulation schemes have appeared in [20], [21] and [14].

The switched network model considered here requires the arrival processes to be Poisson. However, this is not a major restriction, due to a *Poissonization* trick considered, for example in [16] and [30]: all arriving packets are first passed through a "regularizer", which emits packets according to a Poisson process with a rate that lies between the arrival rate and the network capacity. This leads to the arrivals being effectively Poisson, as seen by the system, with a somewhat higher rate — by choosing the rate of "regularizer" so that the effective gap to the capacity, i.e., $(1-\rho)$, is decreased by factor 2.

The scheduling policy that we propose is not optimal for general switched networks. For example, in the context of ad hoc wireless networks, in the independent-set model, the number of constraints is equal to the number of edges in the interference graph, which is often much larger than the number of nodes. Under our policy, the average total queue size would scale with the number of edges, whereas maximum-weight policy achieves a scaling proportional to the number of nodes.

There are many possible directions for future research. One direction that is close to the results in this chapter concerns the extension of **EMUL** to multi-hop switched networks. A natural attempt would be to consider a continuous-time analog of a multi-hop switched network, identify an insensitive rate allocation policy with good performance properties, and emulate this continuous-time policy. However, there are various technical difficulties with this approach, and it is not entirely clear how to resolve them.

Another direction is the search for low-complexity scheduling policies with optimal

performance. In the context of input-queued switches, our policy has a complexity that is exponential in N, the number of queues, because one has to compute the sum of exponentially many terms at every time instance. This begs the question of finding an optimal policy with polynomial complexity in N. One candidate is the class of MW-α policies, which has polynomial complexity, but its optimality appears difficult to analyze. Another possible candidate could be, as discussed in the introduction, and in Section 4.8, Chapter 4, a randomized version of proportional fairness. The relationship between **SFA** and proportional fairness is explored in [63], and indeed, in a certain sense that can be made precise, **SFA** converges to proportional fairness. The question remains whether (a version of) proportional fairness is optimal for input-queued switches (cf. the discussion in Section 4.7 as well).

A third interesting direction to pursue has to do with the analysis of different limiting regimes. We are interested in two limits: $N \to \infty$, and $\rho \to 1$, where N is the number of queues, and ρ is the system load. Again, take the example of input-queued switches. In this chapter, we have considered the heavy-traffic limit, i.e., $\rho \to 1$, and show that our policy is optimal. However, if we take the limit $N \to \infty$, while keeping ρ fixed, then the average total queue size scales as $N^{3/2}$, whereas the maximum-weight policy produces a bound of N. A more interesting question relates to the regime where $(1-\rho)\sqrt{N}$ remain bounded, and where $N \to \infty$. In this regime, under either our policy, the maximum-weight policy, or a batching policy in [44], the average total queue sizes scale as $O(N^{3/2})$. We will see in the next chapter that it is possible to break the 3/2 barrier, and achieve an $O(N^{1.25})$ scaling.

Chapter 6

Queue-Size Scaling in Input-Queued Switches

In the previous chapter, we designed a scheduling policy that achieves optimal queue-size scaling in the heavy-traffic regime, in a general $n \times n$ input-queued switch. While this optimality is valid in the regime where the number of ports n is fixed, and the load $\rho \to 1$, a closer inspection reveals that it is also valid whenever $\rho \geq 1 - 1/n^2$. This raises the natural question of determining the optimal queue-size scaling in various other regimes.

In this chapter, we consider the scaling of the long-run average total queue size in an $n \times n$ input-queued switch, in the regime where $\rho = 1 - 1/f(n)$, with $f(n) \geq n$. We focus on the special case where the arrival rates are uniform. The main result of this chapter is a new class of scheduling policies under which the long-run average total queue size scales as $O(n^{1.5} f(n) \log f(n))$. As a corollary, in the regime $f(n) = n$, we obtain an upper bound of order $O(n^{2.5} \log n)$, a substantial improvement upon prior works (where the scaling was $O(n^3)$, ignoring poly-logarithmic dependence on n), including the bound from Chapter 5.

The rest of the chapter is organized as follows. We state our main theorem, Theorem 6.1.1 in Section 6.1. Some preliminaries, which are used in later sections, are introduced in Section 6.2. In Section 6.3, we describe the scheduling policy in detail, followed by its analysis (proof of the main theorem) in Section 6.4. We conclude

the chapter with some discussion on open problems in Section 6.5.

The prerequisite for reading this chapter is the description of the input-queued switch model in Section 2.2.1, Chapter 2.

6.1 Main Theorem

We state our main theorem below.

Theorem 6.1.1 *Consider an $n \times n$ input queued switch, as described in Section 2.2.1, where the arrival processes are independent Bernoulli with uniform arrival rate ρ/n, and $\rho = 1 - 1/f(n)$, with $f(n) \geq n$. Then, there is a scheduling policy under which the average total queue size is upper bounded by $Cn^{1.5} f(n) \log f(n)$, i.e.,*

$$\limsup_{\tau \to \infty} \frac{1}{\tau} \sum_{t=1}^{\tau} \sum_{i,j=1}^{n} Q_{i,j}(t) \leq Cn^{1.5} f(n) \log f(n),$$

where C is a universal constant that does not depend on n.

The following corollary is immediate.

Corollary 6.1.2 *Consider the same setup as in Theorem 6.1.1, where $f(n) = n$. Then there is a scheduling policy under which the average total queue size is upper bounded by $Cn^{2.5} \log n$, i.e.,*

$$\limsup_{\tau \to \infty} \frac{1}{\tau} \sum_{t=1}^{\tau} \sum_{i,j=1}^{n} Q_{i,j}(t) \leq Cn^{2.5} \log n,$$

where C is a universal constant that does not depend on n.

6.2 Preliminaries

Notation and Terminology. Recall the input-queued switch model described in Section 2.2.1, Chapter 2. For a general $n \times n$ input-queued switch, recall that $\mathcal{S}$ denotes the schedule set defined by Eq. (2.10), $\mathbf{Q}(\tau) = [Q_{i,j}(\tau)]_{i,j=1}^{n}$ the queue-size

vector at the beginning of time slot τ, $\mathbf{A}(\tau) = [A_{i,j}(\tau)]_{i,j=1}^{n}$ the number of packets that have arrived up to the beginning of time slot τ, and $\sigma(\tau) \in \mathcal{S}$ the schedule chosen during time slot τ. For each i, j, and $\tau \in N$, we have

$$Q_{i,j}(\tau+1) = Q_{i,j}(\tau) - \sigma_{i,j}(\tau)\mathbf{1}_{\{Q_{i,j}(\tau)>0\}} + A_{i,j}(\tau+1) - A_{i,j}(\tau). \tag{6.1}$$

Without loss of generality, we can assume that $Q_{i,j}(0) = 0$, for all i, j.

We first need the concept of *cumulative effective service.* By summing Eq. (6.1) over time, for each $\tau \in \mathbb{N}$, we get that

$$Q_{i,j}(\tau) = Q_{i,j}(0) + A_{i,j}(\tau) - \sum_{t=0}^{\tau-1} \sigma_{i,j}(t)\mathbf{1}_{\{Q_{i,j}(t)>0\}}. \tag{6.2}$$

If $Q_{i,j}(0) = 0$, and if we write $P_{i,j}(\tau) = \sum_{t=0}^{\tau-1} \sigma_{i,j}(t)\mathbf{1}_{\{Q_{i,j}(t)>0\}}$, then (6.2) reduces to

$$Q_{i,j}(\tau) = A_{i,j}(\tau) - P_{i,j}(\tau).$$

We call $P_{i,j}(\tau)$ the cumulative *effective* service offered to queue (i,j), up to the beginning of time slot τ. Note that $P_{i,j}(\tau)$ is different from $\sum_{t=0}^{\tau-1} \sigma_{i,j}(t)$, the cumulative service offered to queue (i,j), up to time τ.

We also need the following notation. Let $A_{i,j}(\tau,\tau') = A_{i,j}(\tau') - A_{i,j}(\tau)$ be the number of arrivals to queue (i,j) between time slot τ and time slot τ', and let $P_{i,j}(\tau,\tau') = P_{i,j}(\tau') - P_{i,j}(\tau)$ be the total effective service offered to queue (i,j) between time slot τ and time slot τ'.

Concentration Inequalities. We need the following concentration inequalities (Theorem 2.4 in [10]).

Theorem 6.2.1 (Concentration Inequalities) *Let $X_1, X_2, \ldots, X_m$ be independent and identically distributed Bernoulli random variables, with*

$$\mathbb{P}(X_i = 1) = p, \quad \text{and} \quad \mathbb{P}(X_i = 0) = 1 - p,$$

for $i = 1, 2, \ldots, m$. Consider the sum $X = \sum_{i=1}^{m} X_i$, with mean $\mathbb{E}[X] = np$. Then for any $x > 0$, we have

$$\text{(Lower tail)} \qquad \mathbb{P}(X \leq \mathbb{E}[X] - x) \leq \exp\left\{-\frac{x^2}{2\mathbb{E}[X]}\right\}, \tag{6.3}$$

$$\text{(Upper tail)} \qquad \mathbb{P}(X \geq \mathbb{E}[X] + x) \leq \exp\left\{-\frac{x^2}{2(\mathbb{E}[X] + x/3)}\right\}. \tag{6.4}$$

Kingman's Bound for $G/G/1$ Queue. Consider a $G/G/1$ queueing system. More precisely, jobs arrive to the system, requesting service times that are independent and identically distributed (i.i.d) as a random variable L, say. The inter-arrival times of the jobs are also i.i.d as a random variable M, and are independent from the service requirements of incoming jobs. Suppose that we use a First-Come-First-Serve (FCFS) service policy and do not allow pre-emption. Let $\lambda \triangleq 1/\mathbb{E}[L]$ be the arrival rate, $\sigma_t^2 \triangleq \text{Var}(L)$ the variance of inter-arrival times, $\mu \triangleq 1/\mathbb{E}[M]$ the service rate, and $\sigma_x^2 \triangleq \text{Var}(M)$ the variance of the service requirements.

The following bound is well-known.

Theorem 6.2.2 (Kingman's bound) *Consider a $G/G/1$ queueing system under the FCFS policy, with λ, μ, σ_t^2 and σ_x^2 defined earlier. Suppose $\lambda < \mu$. Then, the mean waiting time in queue of a job, W, satisfies*

$$W \leq \frac{\lambda(\sigma_x^2 + \sigma_t^2)}{2(1 - \lambda/\mu)}.$$

Minimum Clearance Time. We also need the concept of *minimum clearance time* of a queue matrix, which can be found in [44]. Consider a certain queue matrix $\mathbf{Q} = (Q_{i,j})_{i,j=1}^{n}$, where $Q_{i,j}$ denotes the number of packets at input port i destined for output port j. Suppose that no new packets enter, and the goal is to simply clear all packets present in the system, in minimum time, using only the feasible schedules/matchings. We call this the *minimum clearance time* of the queue matrix $\mathbf{Q}$, and we denote it by $L(\mathbf{Q})$. Then $L(\mathbf{Q})$ is exactly characterized as follows.

Theorem 6.2.3 *Let $\mathbf{Q} = (Q_{i,j})_{i,j=1}^{n}$ be a queue matrix. Let $R_i = \sum_{j'=1}^{n} Q_{i,j'}$ be the ith row sum, and let $C_j = \sum_{i'=1}^{n} Q_{i',j}$ be the jth column sum. Let $L(\mathbf{Q})$ be the*

minimum clearance time of $\mathbf{Q}$*, defined earlier. Then,*

$$L(\mathbf{Q}) = \max_{i,j} \{R_i, C_j\} \tag{6.5}$$

i.e., the maximum over the row sums and column sums.

6.3 Policy Description

To describe our policy, we first define three parameters, T, S and Y, which specify different lengths of time intervals. They are given by

$$T = 72\big(f(n)\big)^2 \log f(n), \tag{6.6}$$

$$S = 224 n^{0.5} f(n) \log f(n), \tag{6.7}$$

$$Y = \rho T + \sqrt{18 \rho T \log f(n)}, \tag{6.8}$$

respectively. We also define n schedules $\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}, \ldots, \boldsymbol{\pi}^{(n)}$. For $r \in \{1, 2, \ldots, n\}$, $\boldsymbol{\pi}^{(r)}$ is defined by

$$\pi_{i,j}^{(r)} = \begin{cases} 1, & \text{if } j = r + i - 1, \text{ and } i \in \{1, 2, \ldots, n - r + 1\}, \\ 1, & \text{if } j = r + i - 1 - n, \text{ and } i \in \{n - r + 2, \ldots, n\}, \\ 0, & \text{otherwise.} \end{cases}$$

To illustrate, when $n = 3$, the schedules $\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}$ and $\boldsymbol{\pi}^{(3)}$ are given by

$$\boldsymbol{\pi}^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\pi}^{(2)} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\pi}^{(3)} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Note that

$$\boldsymbol{\pi}^{(1)} + \boldsymbol{\pi}^{(2)} + \cdots + \boldsymbol{\pi}^{(n)} = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix},$$

the $n \times n$ matrix of all 1.

We are now in position to describe our policy. We consider both arrivals and schedules in terms of "epochs". An arrival epoch consists of the arrival patterns over a period of T time slots. Thus the first arrival epoch covers time slot 1 through time slot T, and in general, the kth arrival epoch covers time slot $(k-1)T+1$ through time slot kT. The lengths of scheduling epochs are more variable. For any k, scheduling decisions during the kth scheduling epoch are made only based on the kth arrival epoch, and the schedules are dedicated only to arrivals that take place in the kth arrival epoch. We also impose the following constraints.

1. The kth scheduling epoch starts only after all arrivals of the $(k-1)$st arrival epoch have been cleared.

2. The kth scheduling epoch starts only after S time slots have elapsed since the beginning of the kth arrival epoch.

3. For any τ, the schedules during the first τ time slots of the kth scheduling epoch are dedicated only to arrivals that take place in the first $S+\tau$ time slots of the kth arrival epoch.

The detailed scheduling decisions within each scheduling epoch are determined as follows. For the kth scheduling epoch, we use the schedules $\boldsymbol{\pi}^{(1)}$, $\boldsymbol{\pi}^{(2)}$, ..., $\boldsymbol{\pi}^{(n)}$ in a round-robin manner, for the first $T-S$ time slots of this epoch. More precisely, for the first $T-S$ time slots, $\boldsymbol{\pi}^{(r)}$ is used in time slots of the form $nq+r$, where $q \in \mathbb{Z}_+$. We now define a certain queue matrix $\mathbf{Q}^{(k)} = [Q_{i,j}^{(k)}]_{i,j=1}^{n}$. $Q_{i,j}^{(k)}$ is the number of packets waiting to be served at queue (i,j), from the kth arrival epoch, at the end of these $T-S$ time slots. In particular, $Q_{i,j}^{(k)}$ does not account for those packets that arrive after the kth arrival epoch. We then clear these packets in the optimal offline manner, i.e., clear $\mathbf{Q}^{(k)}$ in $L(\mathbf{Q}^{(k)})$ amount of time, as described in Theorem 6.2.3. If $T-S+L(\mathbf{Q}^{(k)}) < Y$, that is, if the scheduling duration described thus far is less than Y, then we use empty schedules for $\left(Y-T+S-L(\mathbf{Q}^{(k)})\right)$ time slots. Note that this ensures that the length of a scheduling epoch is at least Y. This completes the description of a scheduling epoch.

Here we provide some remarks on two features of the scheduling policy described above. First, packets that arrive after the kth arrival epoch are not included in the accounting of $Q_{i,j}^{(k)}$, defined in the previous paragraph. Such an accounting will simplify the analysis of our policy, as it ensures that all schedules in the kth scheduling epoch are devoted to packets that arrive during the kth arrival epoch, so that scheduling epochs become i.i.d random variables. While this accounting may result in efficiency loss, we do not expect the loss to be substantial. Second, the length of a scheduling epoch is required to be at least Y. Again, this may result in efficiency loss, but will simplify the analysis of our policy. Under this condition, the length of a scheduling epoch is highly concentrated around Y, which simplifies the calculation of so-called *epoch delay*, defined in the next section.

Finally, the traditional batching policy, as in [44], corresponds to the case where $S = T$, in our setting.

6.4 Policy Analysis

The analysis of our policy proceeds as follows. We decompose the total queue size in any time slot into the sum of three non-negative terms, and we analyze these three terms separately. Since our policy is described in terms of epochs, we will often index time according to epochs as well. More precisely, let $k \in N$. For $\tau \in \{1, 2, \ldots, T\}$, let $Q_{i,j}^{k}(\tau)$ be the size of queue (i, j) at the beginning of time slot $(k-1)T + \tau$. Here we remark that before the proof of Theorem 6.1.1 in Section 6.4.1, we will analyze the queue sizes $Q_{i,j}^{k}$ within the kth arrival epoch, and for convenience, we will drop the superscript k. One can think of time as being 0 at the start of the kth arrival epoch, so that $Q_{i,j}(\tau)$ is the size of queue (i, j) at time τ. In Section 6.4.1, however, we will bring back the superscript k.

We need the concept of *epoch delay*, defined as follows.

Definition 6.4.1 *Recall the convention that time is* 0 *at the beginning of the kth arrival epoch. Suppose that the starting time of the kth scheduling epoch is* $S + D$*; that is,* $S + D$ *time slots after the start of the kth arrival epoch. Then* D *is the delay*

incurred due to previous scheduling epochs, and is called the epoch delay *associated with the kth scheduling epoch.*

When the context is clear, we simply call D the epoch delay. Note that $D \geq 0$, because by Constraint 2 of Section 6.3, the kth scheduling epoch starts only after S time slots have elapsed since the beginning of the kth arrival epoch.

$Q_{i,j}(\tau)$ can be decomposed as the sum of the following three terms:

(i) $Q_{i,j}^{\text{IDEAL}}(\tau - D)$: "ideal" queue size at time $\tau - D$ as if there were no "delay" incurred from previous scheduling epochs;

(ii) $A_{i,j}(\tau - D, \tau)$: extraneous arrivals due to the epoch delay D; and

(iii) $O_{i,j}(\tau)$: leftover packets from the previous arrival epoch.

We now explain this decomposition and the meanings of these three terms in more detail. Suppose for now that $\tau > S + D$ (refer to Figure 6-1). Then $A_{i,j}(\tau - D)$ is the total number of arrivals to queue (i, j) up to time $\tau - D$, and $P_{i,j}(S + D, \tau)$ is the total *effective* service offered to the arrivals in the first $\tau - D$ time slots. Hence,

$$Q_{i,j}^{\text{IDEAL}}(\tau - D) = A_{i,j}(\tau - D) - P_{i,j}(S + D, \tau).$$

$Q_{i,j}^{\text{IDEAL}}(\tau - D)$ is called the "ideal" queue size, because it is precisely the size of queue (i, j) at time $\tau - D$, if the kth scheduling epoch starts precisely at time S. This is further illustrated in Figure 6-2. In the actual system, there is a delay of length D, and hence $A_{i,j}(\tau - D, \tau)$ (highlighted in red in Figure 6-1) is the number of extra arrivals within a period of D time slots. To summarize,

$$Q_{i,j}(\tau) = A_{i,j}(\tau - D) + A_{i,j}(\tau - D, \tau) - P_{i,j}(S + D, \tau) = Q_{i,j}^{\text{IDEAL}}(\tau - D) + A_{i,j}(\tau - D, \tau),$$

when $\tau > S + D$. Note that $O_{i,j}(\tau) = 0$, since all packets from a previous arrival epoch had been cleared at time $S + D$.

Now suppose $\tau \leq S + D$. Since packets from the previous arrival epoch may not have been cleared, $O_{i,j}(\tau)$ may be positive. Now consider the new arrivals in the

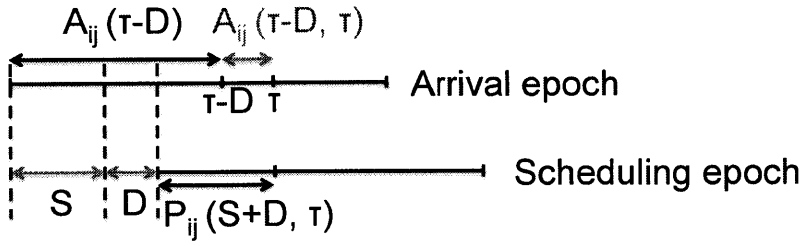


Figure 6-1: Illustration: $\tau > S + D$; actual system

current arrival epoch. Similar to the case where $\tau > S + D$, we consider $A_{i,j}(\tau - D)$, the cumulative number of arrivals up to time $\tau - D$. Note that, however, when $\tau \leq S + D$, we have $\tau + D \leq S$, and there is no service offered to arrivals during the first $\tau - D$ time slots. Hence, in this case,

$$Q_{i.j}^{\text{IDEAL}}(\tau - D) = A_{i,j}(\tau - D).$$

We also need to take into account $A_{i,j}(\tau - D, \tau)$, the number of extra arrivals from time $\tau - D$ to time τ. To summarize,

$$Q_{i,j}(\tau) = O_{i,j}(\tau) + A_{i,j}(\tau - D) + A_{i,j}(\tau - D, \tau) = O_{i,j}(\tau) + Q_{i.j}^{\text{IDEAL}}(\tau - D) + A_{i,j}(\tau - D, \tau).$$

Therefore, in both cases, when $\tau \leq S + D$ and when $\tau > S + D$, we can decompose $Q_{i,j}(\tau)$ as

$$Q_{i,j}(\tau) = O_{i,j}(\tau) + Q_{i.j}^{\text{IDEAL}}(\tau - D) + A_{i,j}(\tau - D, \tau). \tag{6.9}$$

We now analyze the three terms in turn. To do this, we first need to characterize the epoch delay D.

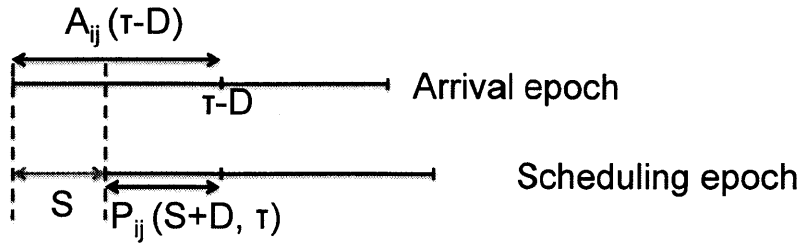


Figure 6-2: Illustration: $\tau > S + D$; ideal system $\mathbf{Q}^{\mathrm{IDEAL}}$

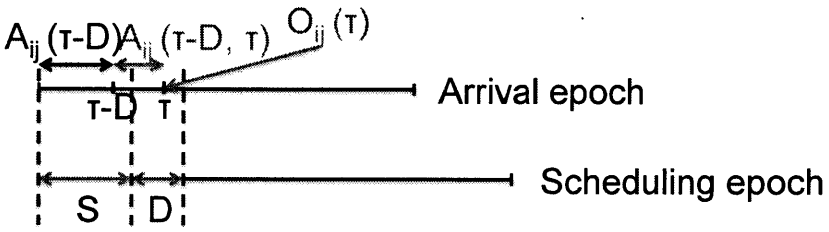


Figure 6-3: Illustration: $\tau \leq S + D$

Analysis of Epoch Delay D

Here we wish to characterize the steady-state epoch delay D. In particular, we want to bound the mean of D in steady state. To do this, we use Kingman's bound (Theorem 6.2.2). As data inputs, we need the mean and variance of the duration of the arrival scheduling epochs. We need some preliminary lemmas that provide bounds for these quantities.

Lemma 6.4.2 *Recall the definition of the quantity Y from Equation (6.8). We have that*

$$Y \leq \frac{1}{2}(1+\rho)T.$$

Proof. Recall that

$$Y = \rho T + \sqrt{18\rho T \log f(n)}.$$

Now

$$\begin{aligned} \frac{1}{2}(1+\rho)T - \rho T &= \frac{1}{2}(1-\rho)T = \frac{1}{2}\left(1 - 1 + \frac{1}{f(n)}\right)T \\ &= \frac{1}{2f(n)} \times 72\big(f(n)\big)^2 \log f(n) = 36 f(n) \log f(n), \end{aligned}$$

and

$$\sqrt{18\rho T \log f(n)} \leq \sqrt{18T \log f(n)} = \sqrt{18 \times 72\big(f(n)\big)^2 \log^2 f(n)} = 36 f(n) \log f(n),$$

thus $\frac{1}{2}(1+\rho)T - \rho T \geq \sqrt{18\rho T \log f(n)}$, and hence

$$Y = \rho T + \sqrt{18\rho T \log f(n)} \leq \frac{1}{2}(1+\rho)T. \qquad \square$$

Lemma 6.4.3 *The length of any scheduling epoch is deterministically upper bounded by*

$$(n+1)T.$$

Proof. Since the arrival processes are Bernoulli, at most one packet arrives to each queue in each time slot. Since n queues are associated with each input port, and with each output port, at most n packets arrive to the same input port, and at most n packets are destined for the same output port, in a time slot. Thus, in T time slots, at most nT packets arrive to any input or output port.

Consider the longest time that is required to clear all packets that arrive during T time slots, under our policy. The queue matrix at the end of the first $T - S$ time slots of the scheduling epoch has row sum (or column sum) at most nT, and hence by Theorem 6.2.3, the clearance time is at most nT. Thus, under our policy, in $(T - S) + nT$ time slots since the start of a scheduling epoch, we must have cleared all packets that arrive during the T time slots, i.e., during the corresponding arrival epoch. Noting that $(T - S) + nT \leq (n+1)T$, we finish the proof of the lemma. $\square$

Lemma 6.4.4 characterizes the mean and variance of the arrival epochs.

Lemma 6.4.4 *The time between the start of consecutive arrival epochs has a mean of T and zero variance.*

Proof. Trivial. $\square$

Lemma 6.4.5 states that in any scheduling epoch, with high probability, the schedules $\boldsymbol{\pi}^{(\ell)}, \ell \in \{1, 2, \ldots, n\}$ (refer to Section 6.3) are never "wasted".

Lemma 6.4.5 *Consider an arbitrary scheduling epoch, say the kth, and let time be 0 at the start of the kth arrival epoch. Let the total number of arrivals to queue (i,j) up to time t be $A_{i,j}(t)$. Define the following event:*

$$E = \left\{ \forall\, \tau \in \{0,1,2,\ldots,T-S\}, \forall\, i,j \in \{1,2,\ldots,n\}, A_{i,j}(S+\tau) \geq \lceil \frac{\tau}{n} \rceil \right\}, \quad (6.10)$$

where $\lceil x \rceil$ is the smaller integer that is no smaller than x. Then

$$\mathbb{P}(E) \geq 1 - 72\big(f(n)\big)^{-5}. \quad (6.11)$$

Proof. First we explain in words what event E means. Consider any fixed queue (i,j). and the first $T-S$ time slots of the current scheduling epoch. Since the schedules $\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}, \ldots, \boldsymbol{\pi}^{(n)}$ (recall their definitions in Section 6.3) are used in a round-robin manner, during any n consecutive time slots, queue (i,j) is served exactly once. Hence, in any consecutive τ time slots, queue (i,j) is served at most $\lceil \frac{\tau}{n} \rceil$ times. Thus, the set E is the event that no queues are ever empty during the first $T-S$ time slots of the current scheduling epoch. The services are never "wasted" in this sense.

We now start the proof of the lemma. For each $i,j \in \{1,2,\ldots,n\}$, and for each $\tau \in \{1,2,\ldots,T-S\}$, consider the event $\left\{ A_{i,j}(S+\tau) < \lceil \frac{\tau}{n} \rceil \right\}$. Since the arrival rate to each queue is uniformly ρ/n,

$$\mathbb{E}[A_{i,j}(S+\tau)] = \frac{\rho}{n}(S+\tau).$$

By Inequality (6.3) of Theorem 6.2.1, for any $x > 0$,

$$\mathbb{P}\left(A_{i,j}(S+\tau) \leq \frac{\rho}{n}(S+\tau) - x \right) \leq \exp\left\{ -\frac{x^2}{2\rho(S+\tau)/n} \right\}.$$

Let $x = \sqrt{20\frac{\rho}{n}(S+\tau)\log f(n)}$. Then,

$$\mathbb{P}\left(A_{i,j}(S+\tau) \leq \frac{\rho}{n}(S+\tau) - \sqrt{20\frac{\rho}{n}(S+\tau)\log f(n)} \right) \;\leq\; \exp\left\{ -\frac{x^2}{2\rho(S+\tau)/n} \right\}$$

$$= \exp\left\{-\frac{20\rho(S+\tau)\log f(n)/n}{2\rho(S+\tau)/n}\right\}$$
$$= \big(f(n)\big)^{-10}.$$

Furthermore, for all $\tau \in \{1, 2, \ldots, T-S\}$,

$$\lceil \frac{\tau}{n} \rceil \leq \frac{\rho}{n}(S+\tau) - \sqrt{20\frac{\rho}{n}(S+\tau)\log f(n)}.$$

To see this, consider $\sqrt{20\frac{\rho}{n}(S+\tau)\log f(n)}$. For all $\tau \in \{0, 1, 2, \ldots, T-S\}$,

$$\begin{aligned}\sqrt{20\frac{\rho}{n}(S+\tau)\log f(n)} &\leq \sqrt{20\frac{\rho}{n}T\log f(n)} \\ &\leq \sqrt{20\frac{1}{n}\Big(72\big(f(n)\big)^2\log f(n)\Big)\log f(n)} \\ &= \sqrt{20\times 72}n^{-0.5}f(n)\log f(n) \leq 39n^{-0.5}f(n)\log f(n).\end{aligned}$$

On the other hand, consider $\frac{\rho}{n}(S+\tau) - \lceil \frac{\tau}{n} \rceil$. We consider the case where $n \geq 2$, so that $\rho = 1 - 1/f(n) \geq 1 - 1/n \geq 1/2$. For all $\tau \in \{0, 1, 2, \ldots, T-S\}$, we have

$$\begin{aligned}\frac{\rho}{n}(S+\tau) - \lceil \frac{\tau}{n} \rceil &\geq \frac{\rho}{n}(S+\tau) - \frac{\tau}{n} - 1 = \frac{\rho}{n}S - \frac{(1-\rho)\tau}{n} - 1 \\ &\geq \frac{\rho}{n}S - \frac{1-\rho}{n}T - 1 = \frac{\rho}{n}S - \frac{T}{nf(n)} - 1 \\ &\geq \frac{1}{2n}\cdot\big(224n^{0.5}f(n)\log f(n)\big) - \frac{72\big(f(n)\big)^2\log f(n)}{nf(n)} - 1 \\ &= 113n^{-0.5}f(n)\log f(n) - 72n^{-1}f(n)\log f(n) - 1.\end{aligned}$$

Now for $n \geq 2$,

$$\begin{aligned}&\big(112n^{-0.5}f(n)\log f(n) - 72n^{-1}f(n)\log f(n) - 1\big) - 39n^{-0.5}f(n)\log f(n) \\ &= 73n^{-0.5}f(n)\log f(n) - 72n^{-1}f(n)\log f(n) - 1 \geq 0,\end{aligned}$$

and so for $n \geq 2$, and for all $\tau \in \{0, 1, 2, \ldots, T-S\}$,

$$\sqrt{20\frac{\rho}{n}(S+\tau)\log f(n)} \leq 39n^{-0.5}f(n)\log f(n)$$

$$\begin{aligned} &\leq 112n^{-0.5}f(n)\log f(n) - 72n^{-1}f(n)\log f(n) - 1 \\ &\leq \frac{\rho}{n}(S+\tau) - \lceil\frac{\tau}{n}\rceil, \end{aligned}$$

and hence

$$\lceil\frac{\tau}{n}\rceil \leq \frac{\rho}{n}(S+\tau) - \sqrt{20\frac{\rho}{n}(S+\tau)\log f(n)}.$$

In summary, for any i, j, for any $\tau \in \{0, 1, 2, \ldots, T-S\}$,

$$\begin{aligned} \mathbb{P}\left(A_{i,j}(S+\tau) < \lceil\frac{\tau}{n}\rceil\right) &\leq \mathbb{P}\left(A_{i,j}(S+\tau) \leq \frac{\rho}{n}(S+\tau) - \sqrt{20\frac{\rho}{n}(S+\tau)\log f(n)}\right) \\ &\leq \big(f(n)\big)^{-10}. \end{aligned}$$

Using the union bound, we have

$$\begin{aligned} &\mathbb{P}\left(\exists\, \tau \in \{1,2,\ldots,T-S\}, \exists\, i,j \in \{1,2,\ldots,n\} \text{ such that } A_{i,j}(S+\tau) < \lceil\frac{\tau}{n}\rceil\right) \\ &\leq \sum_{\tau=0}^{T-S}\sum_{i,j=1}^{n} \mathbb{P}\left(A_{i,j}(S+\tau) < \lceil\frac{\tau}{n}\rceil\right) \leq \sum_{\tau=0}^{T-S}\sum_{i,j=1}^{n} \big(f(n)\big)^{-10} \\ &\leq Tn^2\big(f(n)\big)^{-10} = 72\big(nf(n)\big)^2\big(f(n)\big)^{-10}\log f(n) \\ &\leq 72\big(f(n)\big)^5\big(f(n)\big)^{-10} = 72\big(f(n)\big)^{-5}. \end{aligned}$$

Noting that the complement of event E, say E^c, is given by

$$E^c = \left\{\exists\, \tau \in \{1,2,\ldots,T-S\}, \exists\, i,j \in \{1,2,\ldots,n\} \text{ such that } A_{i,j}(S+\tau) < \lceil\frac{\tau}{n}\rceil\right\},$$

we have

$$\mathbb{P}(E) \geq 1 - 72\big(f(n)\big)^{-5}.$$

□

Lemma 6.4.6 states that the minimum clearance time of all arrivals in an arrival epoch is upper bounded by Y (recall Equation (6.8)), with high probability.

Lemma 6.4.6 *Consider an arbitrary arrival epoch, say the kth, and let the total number of arrivals to queue (i,j) during this epoch be $A_{i,j}$. Let $R_i = \sum_{j'=1}^{n} A_{i,j'}$,*

$C_j = \sum_{i'=1}^{n} A_{i',j}$, *and* $L = \max_{i,j}\{R_i, C_j\}$. *Then*

$$\mathbb{P}(L \leq Y) \geq 1 - 2\big(f(n)\big)^{-5}. \tag{6.12}$$

Proof. Let $i \in \{1, 2, \ldots, n\}$. Then

$$\mathbb{E}[R_i] = \mathbb{E}\left[\sum_{j'=1}^{n} A_{i,j'}\right] = \sum_{j'=1}^{n} \mathbb{E}[A_{i,j'}] = \sum_{j'=1}^{n} \frac{\rho}{n} T = \rho T.$$

By Inequality (6.4) of Theorem 6.2.1, for any $x > 0$,

$$\mathbb{P}(R_i \geq \rho T + x) \leq \exp\left\{-\frac{x^2}{2(\rho T + x/3)}\right\}.$$

Let $x = \sqrt{18\rho T \log f(n)}$, so that $\rho T + x = Y$ (recall Equation (6.8)). Thus, we have that

$$\begin{aligned}
\mathbb{P}(R_i \geq Y) &= \mathbb{P}\left(R_i \geq \rho T + \sqrt{18\rho T \log f(n)}\right) \\
&\leq \exp\left\{-\frac{18\rho T \log f(n)}{2(\rho T + x/3)}\right\} \leq \exp\left\{-\frac{18\rho T \log f(n)}{2(\rho T + x)}\right\} \\
&= \exp\left\{-\frac{9\rho T \log f(n)}{Y}\right\} \leq \exp\left\{-\frac{9\rho T \log f(n)}{\frac{1}{2}(1+\rho)T}\right\} \text{ (Lemma 6.4.2)} \\
&= \exp\left\{-\frac{18\rho \log f(n)}{(1+\rho)}\right\}.
\end{aligned}$$

If $n \geq 2$, then $\rho = 1 - 1/f(n) \geq 1 - 1/n \geq 1/2$, and hence $\rho/(1+\rho) \geq 1/3$. In this case,

$$\mathbb{P}(R_i \geq Y) \leq \exp\left\{-\frac{18\rho \log f(n)}{(1+\rho)}\right\} \leq \exp\big(-6\log f(n)\big) = \big(f(n)\big)^{-6}.$$

By symmetry, we have similar bounds for the C_j:

$$\mathbb{P}(C_j \geq Y) \leq \big(f(n)\big)^{-6}.$$

Since $L = \max_{i,j}\{R_i, C_j\}$,

$$\begin{aligned}
\mathbb{P}(L \geq Y) &= \mathbb{P}\left(\max_{i,j}\{R_i, C_j\} \geq Y\right) \\
&= \mathbb{P}(\exists\ i \text{ s.t. } R_i \geq Y \text{ or } \exists\ j \text{ s.t. } C_j \geq Y) \\
&\leq \sum_{i=1}^{n} \mathbb{P}(R_i \geq Y) + \sum_{j=1}^{n} \mathbb{P}(C_j \geq Y) \\
&\leq \sum_{i=1}^{n} \big(f(n)\big)^{-6} + \sum_{j=1}^{n} \big(f(n)\big)^{-6} \leq 2\big(f(n)\big)^{-5},
\end{aligned}$$

where we used the fact that $f(n) \geq n$. Thus,

$$\mathbb{P}(L \leq Y) \geq 1 - \mathbb{P}(L \geq Y) \geq 1 - 2\big(f(n)\big)^{-5}. \qquad \square$$

Lemmas 6.4.5 and 6.4.6 together imply that the length of a scheduling epoch is with high probability upper bounded by Y.

Corollary 6.4.7 *Let Z be the random variable representing the length of a scheduling epoch. Then*

$$\mathbb{P}(Z \leq Y) \geq 1 - 74\big(f(n)\big)^{-5}. \tag{6.13}$$

Proof. We carry notation from Lemma 6.4.5 and 6.4.6. More specifically, let time be 0 at the start of an arrival epoch. let $A_{i,j}(t)$ denote the number of packets that have arrived to queue (i,j) up to time t, and let $A_{i,j}$ be the total number of packets that have arrived to queue (i,j) during the entire arrival epoch. Let the event E be defined as in (6.10), and let L be defined as in Lemma 6.4.6. Then consider the event $E \cap \{L \leq Y\}$. Since none of the schedules $\boldsymbol{\pi}^{(1)}, \ldots, \boldsymbol{\pi}^{(n)}$ was ever wasted under event E, it was as if we were scheduling the packets in the optimal offline manner, i.e., having collected all packets that have arrived in the entire arrival epoch, and serving them in the optimal manner (see Theorem 6.2.3). Therefore, under event E, $Z = L$, and so

$$E \cap \{L \leq Y\} \subset \{Z \leq Y\}.$$

Thus,

$$\begin{aligned}
\mathbb{P}(Z \leq Y) &\geq \mathbb{P}(E \cap \{L \leq Y\}) \\
&= 1 - \mathbb{P}\left(\left(E \cap \{L \leq Y\}\right)^c\right) \\
&= 1 - \mathbb{P}\left(E^c \cup \{L \leq Y\}^c\right) \\
&\geq 1 - \mathbb{P}(E^c) - \mathbb{P}(\{L \leq Y\}^c) \\
&= 1 - [1 - \mathbb{P}(E)] - [1 - \mathbb{P}(L \leq Y)] \\
&\geq 1 - 72\big(f(n)\big)^{-5} - 2\big(f(n)\big)^{-5} \\
&= 1 - 74\big(f(n)\big)^{-5},
\end{aligned}$$

where the last inequality follows from Lemma 6.4.5 and 6.4.6. □

Mean and Variance of a Scheduling Epoch.

Corollary 6.4.8 *As in Corollary 6.4.7, let Z be the random variable representing the length of a scheduling epoch. Let $\mathbb{E}[Z]$ be its mean, and let $Var(Z)$ be its variance. Then,*

$$\mathbb{E}[Z] \leq \frac{1}{2}(1+\rho)T + C_1, \text{ and } Var(Z) \leq C_2 T,$$

for some universal constants C_1 and C_2.

Proof. By the law of total expectation, we have that

$$\mathbb{E}[Z] = \mathbb{E}[Z \mid Z \leq Y]\,\mathbb{P}(Z \leq Y) + \mathbb{E}[Z \mid Z > Y]\,\mathbb{P}(Z > Y). \tag{6.14}$$

For the first term on the RHS of Equation (6.14), we have that

$$\mathbb{E}[Z \mid Z \leq Y]\,\mathbb{P}(Z \leq Y) \leq Y.$$

For the second term on the RHS of Equation (6.14), by Corollary 6.4.7 and Lemma

6.4.3, we have that

$$\begin{aligned}\mathbb{E}\left[Z \mid Z > Y\right]\mathbb{P}\left(Z > Y\right) &\leq (n+1)T\left(1-\mathbb{P}(Z \leq Y)\right) \text{ (Lemma 6.4.3)} \\ &\leq (n+1)\left(72\big(f(n)\big)^2 \log f(n)\right)\left(74\big(f(n)\big)^{-5}\right) \text{ (Corollary 6.4.7)} \\ &\leq C_1,\end{aligned}$$

for some universal constant C_1.

By Lemma 6.4.2, $Y \leq \frac{1}{2}(1+\rho)T$. Thus, in summary, we have that

$$\mathbb{E}[Z] \leq Y + C_1 \leq \frac{1}{2}(1+\rho)T + C_1.$$

We now compute bounds for $\mathrm{Var}(Z)$. Using $\mathrm{Var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}^2[Z]$, we consider $\mathbb{E}[Z^2]$ and $\mathbb{E}^2[Z]$ respectively.

First, note that by construction, we always have $Z \geq Y$. Thus, $\mathbb{E}^2[Z] \geq Y^2$. Second, using again the law of total expectation,

$$\mathbb{E}[Z^2] = \mathbb{E}\left[Z^2 \mid Z \leq Y\right]\mathbb{P}\left(Z \leq Y\right) + \mathbb{E}\left[Z^2 \mid Z > Y\right]\mathbb{P}\left(Z > Y\right). \tag{6.15}$$

For the first term on the RHS of Equation (6.15), we have that

$$\mathbb{E}\left[Z^2 \mid Z \leq Y\right]\mathbb{P}\left(Z \leq Y\right) \leq Y^2.$$

For the second term on the RHS of Equation (6.15), again, by Corollary 6.4.7 and Lemma 6.4.3, we have that

$$\begin{aligned}\mathbb{E}\left[Z^2 \mid Z > Y\right]\mathbb{P}\left(Z > Y\right) &\leq (n+1)^2T^2\left(1-\mathbb{P}(Z \leq Y)\right) \\ &\leq (n+1)^2\left(72\big(f(n)\big)^2 \log f(n)\right)T\left(74\big(f(n)\big)^{-5}\right) \\ &\leq C_2T,\end{aligned}$$

for some universal constant C_2. Thus, we have that

$$\mathbb{E}[Z^2] \leq Y^2 + C_2 T,$$

and so

$$\mathrm{Var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}^2[Z] \leq Y^2 + C_2 T - Y^2 = C_2 T. \qquad \square$$

Mean of Epoch Delay D in Steady State.

Proposition 6.4.9 *There exists universal constants C_2 and C_3 such that in steady state,*

$$\mathbb{E}[D] \leq C_2 f(n) + C_3.$$

Proof. To bound the steady-state mean of epoch delay D, we use Kingman's bound (Theorem 6.2.2). Let λ be the arrival rate of the arrival epochs, and let σ_t^2 be the variance. Then, by Lemma 6.4.4,

$$\lambda = \frac{1}{T}, \quad \text{and} \quad \sigma_t^2 = 0.$$

Let μ be the service rate of the scheduling epochs, i.e., $\mu = 1/\mathbb{E}[Z]$, where Z is the random variable that denotes the length of a scheduling epoch. We also let $\sigma_x^2 = \mathrm{Var}(Z)$ be the variance of Z. Then by Corollary 6.4.8,

$$\mu \geq \frac{1}{(1+\rho)T/2 + C_1}, \quad \text{and} \quad \sigma_x^2 \leq C_2 T,$$

for some universal constants C_1 and C_2. Hence,

$$\frac{\lambda}{\mu} \leq \frac{C_1 + (1+\rho)T/2}{T}.$$

By Kingman's bound,

$$\begin{aligned}
\mathbb{E}[D] &\leq \frac{\lambda(\sigma_x^2+\sigma_t^2)}{2(1-\lambda/\mu)} \leq \frac{1}{T}\cdot\frac{C_2T}{2\left(1-\frac{C_1+(1+\rho)T/2}{T}\right)} \\
&= \frac{C_2T}{2\left(\frac{1-\rho}{2}T-C_1\right)} = \frac{C_2T}{(1-\rho)\left(T-\frac{2C_1}{1-\rho}\right)} \\
&= \frac{C_2\left(T-\frac{2C_1}{1-\rho}\right)+\frac{2C_1C_2}{1-\rho}}{(1-\rho)\left(T-\frac{2C_1}{1-\rho}\right)} \\
&= \frac{C_2}{1-\rho}+\frac{2C_1C_2}{(1-\rho)^2T-2C_1(1-\rho)} \\
&= \frac{C_2}{1-\rho}+C_3 = C_2f(n)+C_3,
\end{aligned}$$

for some universal constant C_3. □

Analysis of $Q_{i,j}^{\text{IDEAL}}$

We now analyze $Q_{i,j}^{\text{IDEAL}}$. Let $i,j \in \{1,2,\ldots,n\}$, let $k \in N$, and let time be 0 at the start of the kth arrival epoch. Let $\tau \in \{1,2,\ldots,T\}$. Recall that $Q_{i,j}^{\text{IDEAL}}(\tau)$ is the size of queue (i,j) at time τ, when the epoch delay $D=0$. If $\tau > S$, as illustrated in Figure 6-4, then

$$Q_{i,j}^{\text{IDEAL}}(\tau) = A_{i,j}(\tau) - P_{i,j}(S,\tau),$$

where $A_{i,j}(\tau)$ is the cumulative number of arrivals up to time τ, and $P_{i,j}(S,\tau)$ is the cumulative effective service offered to these arrivals. If $\tau \leq S$, as illustrated in Figure 6-5, then

$$Q_{i,j}^{\text{IDEAL}}(\tau) = A_{i,j}(\tau),$$

since $P_{i,j}(S,\tau)=0$. We now wish to bound $\mathbb{E}\left[Q_{i,j}^{\text{IDEAL}}(\tau)\right]$.

Proposition 6.4.10 *Let $k \in N$, and let time be 0 at the start of the kth arrival epoch. Then for any $\tau \in \{1,2,\ldots,T\}$,*

$$\mathbb{E}\left[Q_{i,j}^{IDEAL}(\tau)\right] \leq 224n^{-0.5}f(n)\log f(n) + C_4, \tag{6.16}$$

for some universal constant C_4.

Proof. As per earlier discussions, we consider two cases: when $\tau \leq S$, and when $\tau > S$.

(i) If $\tau \leq S$, then as illustrated in Figure 6-5,

$$Q_{i,j}^{\text{IDEAL}}(\tau) = A_{i,j}(\tau),$$

since by Constraint 2 of Section 6.3, $P_{i,j}(S,\tau) = 0$ for all $\tau \leq S$. Thus

$$\begin{aligned} \mathbb{E}\left[Q_{i,j}^{\text{IDEAL}}(\tau)\right] &= \mathbb{E}\left[A_{i,j}(\tau)\right] \leq \mathbb{E}\left[A_{i,j}(S)\right] \\ &\leq \frac{\rho}{n} S \leq \frac{1}{n} S = 224 n^{-0.5} f(n) \log f(n). \end{aligned}$$

(ii) If $\tau > S$, then as illustrated in Figure 6-4,

$$Q_{i,j}^{\text{IDEAL}}(\tau) = A_{i,j}(\tau) - P_{i,j}(S,\tau).$$

First we have

$$\mathbb{E}\left[A_{i,j}(\tau)\right] = \frac{\rho}{n}\tau \leq \frac{\tau}{n} = \frac{S}{n} + \frac{\tau - S}{n}.$$

Next we show that

$$\mathbb{E}\left[P_{i,j}(S,\tau)\right] \geq \frac{\tau - S}{n} - C_4,$$

for some universal constant C_4. To do this, recall the definition (6.10) of event E from Lemma 6.4.5:

$$E = \left\{ \forall\, \tau' \in \{0,1,2,\ldots,T-S\}, \forall\, i,j \in \{1,2,\ldots,n\}, A_{i,j}(S+\tau') \geq \lceil \frac{\tau'}{n} \rceil \right\}.$$

Under this event, $P_{i,j}(S,\tau) \geq \lceil \frac{\tau - S}{n} \rceil$. By Lemma 6.4.5, we also have $\mathbb{P}(E) \geq$

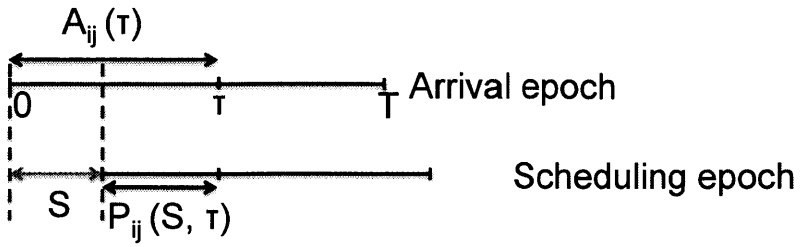


Figure 6-4: Epoch delay $D = 0$; $\tau > S$.

$1 - 72\big(f(n)\big)^{-5}$, and hence

$$\begin{aligned}
\mathbb{E}\left[P_{i,j}(S,\tau)\right] &= \mathbb{E}\left[P_{i,j}(S,\tau) \mid E\right]\mathbb{P}(E) + \mathbb{E}\left[P_{i,j}(S,\tau) \mid E^c\right]\mathbb{P}(E^c) \\
&\geq \lceil \frac{\tau - S}{n} \rceil \left(1 - 72\big(f(n)\big)^{-5}\right) \\
&\geq \frac{\tau - S}{n} - C_4,
\end{aligned}$$

for some universal constant C_4.

In summary,

$$\begin{aligned}
\mathbb{E}\left[Q_{i,j}^{\text{IDEAL}}(\tau)\right] &= \mathbb{E}\left[A_{i,j}(\tau)\right] - \mathbb{E}\left[P_{i,j}(S,\tau)\right] \\
&\leq \frac{S}{n} + \frac{\tau - S}{n} - \left(\frac{\tau - S}{n} - C_4\right) \\
&= \frac{S}{n} + C_4 = 224 n^{-0.5} f(n) \log f(n) + C_4.
\end{aligned}$$

Since under both cases $\tau \leq S$ and $\tau > S$,

$$\mathbb{E}\left[Q_{i,j}^{\text{IDEAL}}(\tau)\right] \leq 224 n^{-0.5} f(n) \log f(n) + C_4,$$

we have established (6.21). □

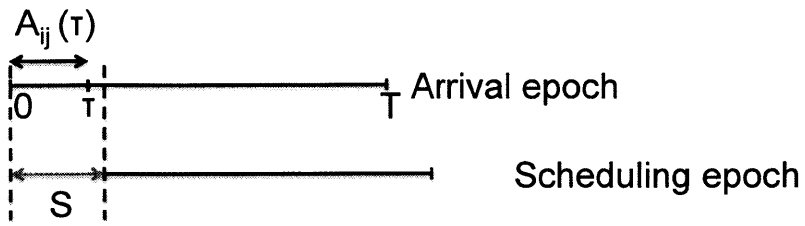


Figure 6-5: Epoch delay $D = 0$; $\tau \leq S$.

Analysis of $O_{i,j}(\tau)$ and $A_{i,j}(\tau - D, \tau)$

Let $i, j \in \{1, 2, \ldots, n\}$, let $k \in N$, and let time be 0 at the start of the kth arrival epoch. Let $\tau \in \{1, 2, \ldots, T\}$. Recall the decomposition (6.9) of $Q_{i,j}(\tau)$:

$$Q_{i,j}(\tau) = O_{i,j}(\tau) + Q_{i,j}^{\text{IDEAL}}(\tau - D) + A_{i,j}(\tau - D, \tau).$$

Here $O_{i,j}(\tau)$ is the number of leftover packets from the previous arrival epoch, and $A_{i,j}(\tau - D, \tau)$ is the number of extraneous arrivals due to the epoch delay D. We now characterize $O_{i,j}(\tau)$ and $A_{i,j}(\tau - D, \tau)$ in terms of D.

Lemma 6.4.11 *Let $O_{i,j}$ be defined as earlier. If the epoch delay is D, then for any $\tau \in \{1, 2, \ldots, T\}$,*

$$\sum_{i,j=1}^{n} O_{i,j}(\tau) \leq n(S + D). \tag{6.17}$$

Proof. Recall that by our policy, at time $S + D$, all packets from a previous arrival epoch have been cleared. Thus for $\tau > S + D$, $O_{i,j}(\tau) = 0$. For $\tau \leq S + D$, since the total amount of service provided to all queues over a time period of length $S + D$ cannot exceed $n(S + D)$, and all packets from a previous arrival epoch have been cleared at time $S + D$, we have that

$$\sum_{i,j=1}^{n} O_{i,j}(\tau) \leq n(S + D).$$

This concludes the proof. □

Lemma 6.4.12 *Let $\tau \in \{1, 2, \ldots, T\}$, and let the epoch delay be D. Let $A_{i,j}(\tau - D, \tau)$ be defined as earlier. Then,*

$$\mathbb{E}\left[A_{i,j}(\tau - D, \tau) \mid D\right] \leq \frac{\rho}{n} D. \tag{6.18}$$

Proof. For any $t \in N$, let $a_{i,j}(t)$ be the number of arrivals during time slot t. Suppose for now that $\tau \geq D$. Then

$$A_{i,j}(\tau - D, \tau) = \sum_{t=\tau-D+1}^{\tau} a_{i,j}(t).$$

Since the epoch delay D is a random variable that is determined by previous arrival epochs, and arrivals are independent across time slots, $a_{i,j}(t)$ is independent from D, for $t \in \{\tau - D + 1, \tau\}$. Thus

$$\begin{aligned}
\mathbb{E}\left[A_{i,j}(\tau - D, \tau) \mid D\right] &= \mathbb{E}\left[\sum_{t=\tau-D+1}^{\tau} a_{i,j}(t) \mid D\right] \\
&= \sum_{t=\tau-D+1}^{\tau} \mathbb{E}\left[a_{i,j}(t)\right] = \sum_{t=\tau-D+1}^{\tau} \frac{\rho}{n} \\
&= \frac{\rho}{n} D.
\end{aligned}$$

If $\tau < D$, then

$$\begin{aligned}
\mathbb{E}\left[A_{i,j}(\tau - D, \tau) \mid D\right] &= \mathbb{E}\left[\sum_{t=1}^{\tau} a_{i,j}(t) \mid D\right] \\
&= \sum_{t=1}^{\tau} \mathbb{E}\left[a_{i,j}(t)\right] = \sum_{t=1}^{\tau} \frac{\rho}{n} \\
&= \frac{\rho}{n} \tau \leq \frac{\rho}{n} D.
\end{aligned}$$

In both cases, we have

$$\mathbb{E}\left[A_{i,j}(\tau - D, \tau) \mid D\right] \leq \frac{\rho}{n} D.$$

This concludes the proof. □

6.4.1 Proof of Main Theorem 6.1.1

We are now ready to prove Theorem 6.1.1. Recall the notation described at the beginning of Section 6.4. Let $k \in N$, and let $\tau \in \{1, 2, \ldots, T\}$. Recall that $Q_{i,j}^k(\tau)$ is the size of queue (i,j) at the beginning of time slot $(k-1)T + \tau$. Let D^k be the epoch delay associated with the kth scheduling epoch. The decomposition (6.9) can be re-written as follows.

$$Q_{i,j}^k(\tau) = O_{i,j}^k(\tau) + Q_{i.j}^{k,\text{IDEAL}}(\tau - D^k) + A_{i,j}^k(\tau - D^k, \tau), \tag{6.19}$$

where

(i) $Q_{i.j}^{k,\text{IDEAL}}(\tau - D^k)$: "ideal" queue size at time $(k-1)T + \tau - D$ as if there were no "delay" incurred from previous scheduling epochs;

(ii) $A_{i,j}^k(\tau - D^k, \tau)$: extraneous arrivals during the kth arrival epoch due to the epoch delay D^k; and

(iii) $O_{i,j}^k(\tau)$: leftover packets from the $(k-1)$st arrival epoch.

Now for any $k \in N$, consider the average total queue size up to time kT, i.e.,

$$\frac{1}{kT} \sum_{\ell=1}^{k} \sum_{\tau=1}^{T} \sum_{i,j=1}^{n} Q_{i,j}^{\ell}(\tau).$$

We can decompose this expression as

$$\begin{aligned}
&\frac{1}{kT} \sum_{\ell=1}^{k} \sum_{\tau=1}^{T} \sum_{i,j=1}^{n} Q_{i,j}^{\ell}(\tau) \\
&= \frac{1}{kT} \sum_{\ell=1}^{k} \sum_{\tau=1}^{T} \sum_{i,j=1}^{n} \left\{ O_{i,j}^{\ell}(\tau) + Q_{i.j}^{\ell,\text{IDEAL}}(\tau - D^{\ell}) + A_{i,j}^{\ell}(\tau - D^{\ell}, \tau) \right\} \\
&= \frac{1}{kT} \sum_{\ell=1}^{k} \sum_{\tau=1}^{T} \sum_{i,j=1}^{n} O_{i,j}^{\ell}(\tau) + \frac{1}{kT} \sum_{\ell=1}^{k} \sum_{\tau=1}^{T} \sum_{i,j=1}^{n} Q_{i.j}^{\ell,\text{IDEAL}}(\tau - D^{\ell}) \\
&\quad + \frac{1}{kT} \sum_{\ell=1}^{k} \sum_{\tau=1}^{T} \sum_{i,j=1}^{n} A_{i,j}^{\ell}(\tau - D^{\ell}, \tau).
\end{aligned}$$

We consider the first term on the RHS first. By Lemma 6.4.11,

$$\begin{aligned}\frac{1}{kT}\sum_{\ell=1}^{k}\sum_{\tau=1}^{T}\sum_{i,j=1}^{n} O_{i,j}^{\ell}(\tau) &\leq \frac{1}{kT}\sum_{\ell=1}^{k}\sum_{\tau=1}^{T}\{n(S+D^{\ell})\} \\ &= nS+\frac{n}{k}\sum_{\ell=1}^{k} D^{\ell}.\end{aligned}$$

Hence,

$$\limsup_{k\to\infty}\sum_{\ell=1}^{k}\sum_{\tau=1}^{T}\sum_{i,j=1}^{n} O_{i,j}^{\ell}(\tau) \leq nS + n\limsup_{k\to\infty}\frac{1}{k}\sum_{\ell=1}^{k} D^{\ell}.$$

By ergodicity and Proposition 6.4.9,

$$\limsup_{k\to\infty}\frac{1}{k}\sum_{\ell=1}^{k} D^{\ell} = \mathbb{E}[D] \leq C_2 f(n) + C_3,$$

for some universal constants C_2 and C_3, and where $\mathbb{E}[D]$ is the steady-state mean of epoch delay D. Thus we have that

$$\limsup_{k\to\infty}\sum_{\ell=1}^{k}\sum_{\tau=1}^{T}\sum_{i,j=1}^{n} O_{i,j}^{\ell}(\tau) \leq nS + nC_2 f(n) + nC_3. \tag{6.20}$$

We now consider the second term

$$\frac{1}{kT}\sum_{\ell=1}^{k}\sum_{\tau=1}^{T}\sum_{i,j=1}^{n} Q_{i.j}^{\ell,\mathrm{IDEAL}}(\tau - D^{\ell}).$$

Note that, first,

$$\sum_{\tau=1}^{T}\sum_{i,j=1}^{n} Q_{i.j}^{\ell,\mathrm{IDEAL}}(\tau - D^{\ell}) \leq \sum_{\tau=1}^{T}\sum_{i,j=1}^{n} Q_{i.j}^{\ell,\mathrm{IDEAL}}(\tau),$$

since the latter term consists of potentially more summands (if $\tau \leq D^{\ell}$, then $Q_{i.j}^{\ell,\mathrm{IDEAL}}\left(\tau - D^{\ell}\right) = 0$). Second, note that

$$\sum_{\tau=1}^{T}\sum_{i,j=1}^{n} Q_{i.j}^{\ell,\mathrm{IDEAL}}(\tau)$$

are independent across different arrival epochs, and are indentically distributed, and hence by the law of large numbers,

$$\frac{1}{k}\sum_{\ell=1}^{k}\sum_{\tau=1}^{T}\sum_{i,j=1}^{n} Q_{i.j}^{\ell,\mathrm{IDEAL}}(\tau) \to \sum_{\tau=1}^{T}\sum_{i,j=1}^{n} \mathbb{E}\left[Q_{i,j}^{\ell,\mathrm{IDEAL}}(\tau)\right]$$

almost surely, as $k \to \infty$. By Proposition 6.4.10,

$$\sum_{\tau=1}^{T}\sum_{i,j=1}^{n} \mathbb{E}\left[Q_{i,j}^{\ell,\mathrm{IDEAL}}(\tau)\right] \leq Tn^2\left(224n^{-0.5}f(n)\log f(n) + C_4\right),$$

for some universal constant C_4. Thus

$$\begin{aligned}&\limsup_{k\to\infty}\frac{1}{kT}\sum_{\ell=1}^{k}\sum_{\tau=1}^{T}\sum_{i,j=1}^{n} Q_{i.j}^{\ell,\mathrm{IDEAL}}(\tau - D^{\ell})\\ &\leq \frac{1}{T}Tn^2\left(224n^{-0.5}f(n)\log f(n) + C_4\right) \leq C_5 n^{1.5} f(n)\log f(n), \qquad (6.21)\end{aligned}$$

for some universal constant C_5.

For the third term

$$\frac{1}{kT}\sum_{\ell=1}^{k}\sum_{\tau=1}^{T}\sum_{i,j=1}^{n} A_{i,j}^{\ell}(\tau - D^{\ell}, \tau),$$

first note that

$$\sum_{\ell=1}^{k}\sum_{\tau=1}^{T}\sum_{i,j=1}^{n} A_{i,j}^{\ell}(\tau - D^{\ell}, \tau) \leq \sum_{\ell=1}^{k}\sum_{\tau=1}^{T}\sum_{i,j=1}^{n} D^{\ell} a_{i,j}^{\ell}(\tau),$$

where $a_{i,j}^{\ell}(\tau)$ is the number of arrivals in time slot $(\ell-1)T+\tau$. The inequality holds because each $a_{i,j}^{\ell}(\tau)$ is summed at most D^{ℓ} times. Thus by ergodicity,

$$\begin{aligned}\limsup_{k\to\infty}\frac{1}{kT}\sum_{\ell=1}^{k}\sum_{\tau=1}^{T}\sum_{i,j=1}^{n} A_{i,j}^{\ell}(\tau - D^{\ell}, \tau) &\leq \limsup_{k\to\infty}\frac{1}{kT}\sum_{\ell=1}^{k}\sum_{\tau=1}^{T}\sum_{i,j=1}^{n} D^{\ell} a_{i,j}^{\ell}(\tau)\\ &= \frac{1}{T}\mathbb{E}[D]\sum_{\tau=1}^{T}\sum_{i,j=1}^{n}\mathbb{E}\left[a_{i,j}^{\ell}(\tau)\right]\\ &= n^2\frac{\rho}{n}\mathbb{E}[D] \leq n\rho\left(C_2 f(n) + C_3\right), \qquad (6.22)\end{aligned}$$

where $\mathbb{E}[D]$ is the steady-state mean of epoch delay D, and C_3 is some universal constant.

Combining Equation (6.20), (6.21) and (6.22), we have that

$$\begin{aligned}
&\limsup_{k\to\infty} \frac{1}{kT} \sum_{\ell=1}^{k} \sum_{\tau=1}^{T} \sum_{i,j=1}^{n} Q_{i,j}^{\ell}(\tau) \\
&\leq nS + nC_2 f(n) + nC_3 + C_5 n^{1.5} f(n) \log f(n) + n\rho \left(C_2 f(n) + C_3\right) \\
&\leq Cn^{1.5} f(n) \log f(n),
\end{aligned}$$

for some universal constant C. By periodicity,

$$\limsup_{k\to\infty} \frac{1}{kT} \sum_{\ell=1}^{k} \sum_{\tau=1}^{T} \sum_{i,j=1}^{n} Q_{i,j}^{\ell} = \limsup_{\tau\to\infty} \frac{1}{\tau} \sum_{t=1}^{\tau} \sum_{i,j=1}^{n} Q_{i,j}(t),$$

and this concludes the proof of Theorem 6.1.1.

6.5 Discussion

We presented a novel scheduling policy for a general $n \times n$ input-queued switch. In the regime where the system load $\rho = 1 - 1/n$, and the arrival rates are uniform, our policy achieves an upper bound of order $O(n^{2.5} \log n)$ on the long-run average total queue size, a substantial improvement upon prior upper bounds, all of which are of order $O(n^3)$, ignoring logarithmic dependence on n.

We now provide several remarks about the policy proposed in this chapter. First, the policy uses detailed knowledge of the arrival statistics, and is heavily dependent on the fact that the arrival rates are uniform. While we believe that similar policies can be devised for arbitrary arrival rates with load $\rho = 1 - 1/f(n)$ $(f(n) \geq n)$, the policy description and analysis are likely to more involved. Second, at a high level, what the policy does is to wait for enough arrivals to take place, so that the system exhibits the desired level of regularity, before it starts any services. This idea itself may be of independent interest. Third and finally, in the regime where $\rho \approx 1 - 1/n$,

we can derive an $\Omega(n^2)$ universal lower bound on the long-run average total queue size (cf. Lemma 5.3.4 in Chapter 5), whereas our upper bound is of order $O(n^{2.5}\log n)$. It is of interest to see whether this gap between the upper and lower bound can be closed. We do not expect an order-improvement on the lower bound; indeed, we believe that there exists an online scheduling policy that achieves an $O(n^2) = O\left(\frac{n}{1-\rho}\right)$ upper bound (ignoring logarithmic dependence on n). First, we have already seen a scheduling policy with an $O\left(\frac{n}{1-\rho}\right)$ upper bound in Chapter 5, when $\rho \geq 1 - 1/n^2$, so it is reasonable to expect this queue-size scaling in other regimes. Second, when the system is in the heavily loaded regime where $\rho \approx 1 - 1/n$, queues are rarely empty, and we expect a good scheduling policy to be able to find full matchings most of the time. The system is then approximately composed of n independent queues, each with load $\rho \approx 1 - 1/n$, and hence an $O(n^2)$ upper bound.

Chapter 7

Concluding Remarks

7.1 Discussion

As detailed contributions of the thesis have already been described in the introduction, Chapter 1, we do not repeat them here. Instead, we provide a quick summary and some additional perspective.

In this thesis, we addressed the design and analysis of various resource allocation schemes in SPNs. We focused on two important instances of SPNs, switched networks and bandwidth-sharing networks. We have primarily considered performance-related questions of resource allocation in these networks. We studied several important performance metrics, including the expected total queue size in the system, in steady state (Chapters 3, 4, 5, and 6), the tail probability of the steady-state queue-size distribution (Chapters 3, 4 and 5), and the maximum excursion of queue sizes over a given time horizon (Chapters 3, and 4).

In Chapters 3 and 4, we made novel uses of Lyapunov function techniques for the study of existing important resource allocation policies (MW-α in switched networks and α-fair in bandwidth-sharing networks), and derived various new insights on performance properties of these policies. In Chapters 5 and 6, we focused on scaling analysis of policies, and designed novel policies with attractive performance measures. A salient feature of the results in this thesis is the explicit dependence of performance bounds on both the network structure, as well as the system load.

We believe that understanding this joint dependence is crucial in designing a desired resource allocation policy.

We now remark on the performance and complexity tradeoff in the context of models and policies considered in this thesis. As mentioned in Chapter 1, two important aspects of a resource allocation policy are performance and the complexity of the calculation required at each step. An ideal policy ought to have low complexity as well as performance guarantees. To illustrate this tradeoff concretely, consider an $n \times n$ input-queued switch, with $N = n^2$ queues. The complexity of a MW policy is well-known to be $O(N^{1.5})$, and for the proportionally fair policy that we proposed in Section 4.7, it is a simple convex program and hence computationally tractable. While both these policies have been conjectured to achieve optimal queue-size scaling in input-queues switches, proving this optimality seems to be beyond the reach of current theory. On the other hand, although the policy **EMUL**, proposed in Chapter 5, achieves provably optimal queue-size scaling in input-queued switches, it has a complexity that is exponential in N. The policy proposed in Chapter 6 is also quite involved, requires information on arrival rates, and applies only to special cases. Therefore, an important open direction is the design (or establishing impossibility) of low-complexity policies with provably good performance, at least in the context of input-queued switches, and more broadly, for more general SPNs. This may involve advancing methods of performance analysis for existing policies, and/or novel insights on how to design good policies.

7.2 Open Problems

At various places in the thesis, we have suggested and discussed open problems/directions for future work. In this section, we list four open problems that are the author's "favorites", and believe that the resolution of one or some of these problems will lead to better understanding of resource allocation policies in SPNs.

1. Extension of EMUL to Multi-Hop Switched Networks. In Chapter 5, we proposed the scheduling policy **EMUL** for single-hop switched networks, which has many attractive performance properties. It is desirable to design a scheduling policy for multi-hop switched networks, with good performance properties as well. One possible direction is to extend the ideas used for the design of **EMUL** to a multi-hop setting. More discussion can be found in Section 5.6.

2. Polynomial-Complexity Approximation of EMUL. The policy **EMUL** from Chapter 5 has a running time that is at least exponential in n, in an $n \times n$ input-queued switch. In general, **EMUL** has a running time that is exponential in N, in a single-hop switched network with N queues. A natural question is to find a polynomial-complexity scheduling policy that reasonably approximates **EMUL**, at least in the context of input-queued switches. Such a policy is likely to have similar performance as **EMUL**. For more discussion, see Section 5.6.

3. Optimal Queue-Size Scaling in Input-Queued Switches with $\rho \approx 1 - 1/n$. Consider an $n \times n$ input-queued switch with load $\rho = 1 - 1/n$. In Chapter 6, we presented a scheduling policy that achieves an $O(n^{2.5} \log n)$ upper bound on the long-run average total queue size, in the case of uniform arrival rates. As discussed in Section 6.5, there is also a universal lower bound of order $\Omega(n^2)$ on the same quantity, so it is of interest to see whether this gap between the upper and lower bound can be closed. We believe that the "right" scaling is $O(n^2)$, so a natural open problem is to design a scheduling policy that achieves an $O(n^2)$ queue-size scaling in an $n \times n$ input-queued switch with $\rho \approx 1 - 1/n$, at least when the arrival rates are uniform.

4. Heavy-Traffic Optimality of Proportional Fairness in Input-Queued Switches. This open problem has appeared as Conjecture 4.7.1, and has been discussed extensively in Section 4.7, so we do not repeat it here.

Appendix A

Proofs Omitted from Chapter 4

A.1 Proof of Lemma 4.6.7

To establish tightness, it suffices to show that for every $y > 0$, there exists a compact set $\mathbb{K}_y \subset \mathbb{R}_+^N$ such that

$$\limsup_{r\to\infty} \boldsymbol{\pi}^r \left(\mathbb{R}_+^N \backslash \mathbb{K}_y\right) \leq e^{-y}. \tag{A.1}$$

We now proceed to define the compact sets $\mathbb{K}_y$. As in the proof of Theorem 4.4.9, let $\varepsilon_r = \varepsilon(\boldsymbol{\lambda}^r)$ be the gap in the rth system. Then, under Assumption 4.4.6, for sufficiently large r, $\varepsilon_r \geq D/r$ for some network-dependent constant $D > 0$. Since $\alpha = 1$, Theorem 4.3.2 implies that for the rth system, there exist load-dependent constants $K_r > 0$ and $\xi_r > 0$ such that for every $\ell \in \mathbb{Z}_+$,

$$\mathbb{P}_{\boldsymbol{\pi}^r}\left(\|\mathbf{M}^r\|_\infty \geq \frac{K_r}{\varepsilon_r} + 2\xi_r\ell\right) \leq \left(\frac{\xi_r}{\xi_r + \varepsilon_r}\right)^{\ell+1}. \tag{A.2}$$

By the definition of a positive load-dependent constant, there exist continuous functions f_1 and f_2 on the open positive orthant such that for all r, $K_r = f_1(\boldsymbol{\mu}^r, \boldsymbol{\nu}^r)$ and $\xi_r = f_2(\boldsymbol{\mu}^r, \boldsymbol{\nu}^r)$. Since $\boldsymbol{\mu}^r \to \boldsymbol{\mu} > \mathbf{0}$ and $\boldsymbol{\nu}^r \to \boldsymbol{\nu} > \mathbf{0}$, we have $K_r \to K \triangleq f_1(\boldsymbol{\mu}, \boldsymbol{\nu}) > 0$ and $\xi_r \to \xi \triangleq f_2(\boldsymbol{\mu}, \boldsymbol{\nu}) > 0$. Define

$$\mathbb{K}_y \triangleq \left\{\mathbf{v} \in \mathbb{R}_+^N : \|\mathbf{v}\|_\infty \leq \frac{(K+1) + 4(\xi+1)^2 \cdot y}{D}\right\}.$$

We now show that (A.1) holds, or equivalently, by the definition of $\mathbb{K}_y$, we show that for every $y > 0$,

$$\limsup_{r\to\infty} \mathbb{P}_{\pi^r}\left(\frac{1}{r}\|\mathbf{M}^r\|_\infty > \frac{(K+1)+4(\xi+1)^2 y}{D}\right) \le e^{-y}. \tag{A.3}$$

Let $\ell_r \triangleq \lfloor 2\xi_r y/\varepsilon_r \rfloor$, where for $z \in \mathrm{R}$, $\lfloor z \rfloor$ is the largest integer not exceeding z. By (A.2), we have

$$\mathbb{P}_{\pi^r}\left(\frac{1}{r}\|\mathbf{M}^r\|_\infty \ge \frac{K_r}{r\varepsilon_r} + \frac{2\xi_r\ell_r}{r}\right) \le \left(\frac{1}{1+\frac{\varepsilon_r}{\xi_r}}\right)^{\ell_r+1}.$$

Taking logarithms on both sides, we have

$$\log \mathbb{P}_{\pi^r}\left(\frac{1}{r}\|\mathbf{M}^r\|_\infty \ge \frac{K_r}{r\varepsilon_r} + \frac{2\xi_r\ell_r}{r}\right) \le -(\ell_r+1)\log\left(1+\frac{\varepsilon_r}{\xi_r}\right).$$

Since $\varepsilon_r \to 0$ and $\xi_r \to \xi > 0$ as $r \to \infty$, $\frac{\varepsilon_r}{\xi_r} < 1$ for sufficiently large r. Since $\log(1+t) \ge t/2$ for $t \in [0,1]$, we have

$$-(\ell_r+1)\log\left(1+\frac{\varepsilon_r}{\xi_r}\right) \le -(\ell_r+1)\frac{\varepsilon_r}{2\xi_r},$$

when r is sufficiently large. By definition, $\ell_r = \lfloor 2\xi_r y/\varepsilon_r \rfloor$, so $\ell_r + 1 \ge 2\xi_r x/\varepsilon_r$, or equivalently, $-(\ell_r+1)\frac{\varepsilon_r}{2\xi_r} \le -y$. Thus, when r is sufficiently large,

$$\log \mathbb{P}_{\pi^r}\left(\frac{1}{r}\|\mathbf{M}^r\|_\infty \ge \frac{K_r}{r\varepsilon_r} + \frac{2\xi_r\ell_r}{r}\right) \le -y.$$

Consider the term $\frac{K_r}{r\varepsilon_r} + \frac{2\xi_r\ell_r}{r}$. When r is sufficiently large, $r\varepsilon_r \ge D$, $K_r \le K+1$, and $\xi_r \le \xi+1$, and so

$$\frac{K_r}{r\varepsilon_r} + \frac{2\xi_r\ell_r}{r} \le \frac{K_r}{r\varepsilon_r} + \frac{2\xi_r(2\xi_r y)}{r\varepsilon_r} \le \frac{K+1}{D} + \frac{4(\xi+1)^2 y}{D}.$$

Thus, for sufficiently large r,

$$\log \mathbb{P}_{\boldsymbol{\pi}^r}\left(\frac{1}{r}\|\mathbf{M}^r\|_\infty > \frac{(K+1)+4(\xi+1)^2 y}{D}\right)$$
$$\leq \log \mathbb{P}_{\boldsymbol{\pi}^r}\left(\frac{1}{r}\|\mathbf{M}^r\|_\infty \geq \frac{K_r}{r\varepsilon_r} + \frac{2\xi_r \ell_r}{r}\right) \leq -y.$$

This establishes (A.3), and also the tightness of $\{\boldsymbol{\pi}^r\}$.

A.2 Proof of Lemma 4.6.8

Next, we prove Lemma 4.6.8. To this end, we need some definitions and background. In particular, we need the concept and properties of *fluid model solutions.*

Definition A.2.1 *A* fluid model solution *(FMS) is an absolutely continuous function* $\mathbf{m} : [0,\infty) \to \mathbb{R}_+^N$ *such that at each regular point* [1] $t > 0$ *of* $\mathbf{m}(\cdot)$*, we have, for each* $i \in \mathcal{I}$*,*

$$\frac{d}{dt} m_i(t) = \begin{cases} \nu_i - \mu_i \phi_i(\mathbf{m}(t)), & \text{if } m_i(t) > 0, \\ 0, & \text{if } m_i(t) = 0, \end{cases} \tag{A.4}$$

and for each $j \in \mathcal{J}$*,*

$$\sum_{i \in \mathcal{I}_+(\mathbf{m}(t))} R_{ji}\phi_i(\mathbf{m}(t)) + \sum_{i \in \mathcal{I}_0(\mathbf{m}(t))} R_{ji}\lambda_i \leq C_j, \tag{A.5}$$

where $\mathcal{I}_+(\mathbf{m}(t)) = \{i \in \mathcal{I} : m_i(t) > 0\}$ *and* $\mathcal{I}_0(\mathbf{m}(t)) = \{i \in \mathcal{I} : m_i(t) = 0\}$*. Note that here* $\mathbf{R}\boldsymbol{\lambda} = \mathbf{C}$*.*

We now collect some properties of a FMS. The following proposition states that the invariant manifold $\mathscr{M}_1$ consists exactly of the stationary points of a FMS.

Proposition A.2.2 (Theorem 4.1 in [32]) *A vector* $\mathbf{m}_0$ *is an invariant state, that is,* $\mathbf{m}_0 \in \mathscr{M}_1$*, if and only if for every fluid model solution* $\mathbf{m}(\cdot)$ *with* $\mathbf{m}(0) = \mathbf{m}_0$*, we have* $\mathbf{m}(t) = \mathbf{m}_0$ *for all* $t > 0$*.*

[1] A point $t \in (0,\infty)$ is a *regular point* of an absolutely continuous function $f : [0,\infty) \to \mathbb{R}_+^N$ if each component of f is differentiable at t. Since $\mathbf{m}$ is absolutely continuous, almost every time $t \in (0,\infty)$ is a regular point for $\mathbf{m}$.

The following theorem states that starting from any initial condition, a FMS will eventually be close to the invariant manifold $\mathscr{M}_1$.

Theorem A.2.3 (Theorem 5.2 in [36]) *Fix $R \in (0,\infty)$ and $\delta > 0$. There is a constant $T_{R,\delta} < \infty$ such that for every fluid model solution $\mathbf{m}(\cdot)$ satisfying $\|\mathbf{m}(0)\|_\infty \leq R$ we have*

$$d(\mathbf{m}(t), \mathscr{M}_1) < \delta, \quad \text{for all } t > T_{R,\delta},$$

where $d(\mathbf{m}(t), \mathscr{M}_1) \triangleq \inf_{\mathbf{m}\in\mathscr{M}_1} \|\mathbf{m} - \mathbf{m}(t)\|_\infty$ is the distance from $\mathbf{m}(t)$ to the manifold $\mathscr{M}_1$.

Proposition A.2.4 states that the value of the Lyapunov function F_1 defined in (4.7) decreases along the path of any FMS.

Proposition A.2.4 (Corollary 6.1 in [36]) *At any regular point t of a fluid model solution $\mathbf{m}(\cdot)$, we have*

$$\frac{d}{dt} F_1(\mathbf{m}(t)) \leq 0,$$

and the inequality is strict if $\mathbf{m}(t) \notin \mathscr{M}_1$.

Using Proposition A.2.4, and the continuity of the lifting map Δ, we can translate Theorem A.2.3 into the following version, which will be used to prove Lemma 4.6.8.

Lemma A.2.5 *Fix $R \in (0,\infty)$ and $\delta > 0$. There is a constant $T_{R,\delta} < \infty$ such that for every fluid model solution $\mathbf{m}(\cdot)$ satisfying $\|\mathbf{m}(0)\|_\infty \leq R$ we have*

$$\|\mathbf{m}(t) - \Delta(\mathbf{w}(t))\|_\infty < \delta, \quad \text{for all } t > T_{R,\delta},$$

where $\mathbf{w}(t) = \mathbf{w}(\mathbf{m}(t))$ is the workload corresponding to $\mathbf{m}(t)$ (see Definition 4.4.7).

Proof. Fix $R > 0$ and $\delta > 0$. Let $\|\mathbf{m}(0)\|_\infty \leq R$. Then,

$$F_1(\mathbf{m}(0)) = \frac{1}{2} \sum_{i\in I} \nu_i^{-1} \kappa_i m_i^2(0) \leq R',$$

where R' depends on R and the system parameters. Since $\mathbf{m}(\cdot)$ is absolutely continuous, by Proposition A.2.4 and the fundamental theorem of calculus, we have that

$F_1(\mathbf{m}(t)) \leq R'$ for all $t \geq 0$. Define the set

$$\mathcal{S} \triangleq \{\mathbf{m} \in \mathbb{R}_+^N : F_1(\mathbf{m}) \leq R'\},$$

and its *δ-fattening*

$$\mathcal{S}_\delta \triangleq \{\mathbf{m} \in \mathbb{R}_+^N : \|\mathbf{m} - \mathbf{m}'\| \leq \delta \text{ for some } \mathbf{m}' \in \mathcal{S}\}.$$

Note that both $\mathcal{S}$ and $\mathcal{S}_\delta$ are compact sets, and $\mathbf{m}(t) \in \mathcal{S} \subset \mathcal{S}_\delta$ for all $t \geq 0$.

Now consider the workload $\mathbf{w}$ defined in Definition 4.4.7. Define the set $\mathbf{w}(\mathcal{S}_\delta) = \{\mathbf{v} \in \mathbb{R}_+^J : \mathbf{v} = \mathbf{w}(\mathbf{m}) \text{ for some } \mathbf{m} \in \mathcal{S}_\delta\}$. Since $\mathbf{w}$ is a linear map, there exists a load-dependent constant H such that

$$\|\mathbf{w}(\mathbf{m}) - \mathbf{w}(\mathbf{m}')\|_\infty \leq H\|\mathbf{m} - \mathbf{m}'\|_\infty,$$

for any $\mathbf{m}, \mathbf{m}' \in \mathbb{R}_+^N$. Thus $\mathbf{w}(\mathcal{S}_\delta)$ is also a compact set. Since $\mathbf{m}(t) \in \mathcal{S}_\delta$ for all $t \geq 0$, $\mathbf{w}(t) \in \mathbf{w}(\mathcal{S}_\delta)$ for all $t \geq 0$. By Proposition 4.6.1, Δ is a continuous map, so Δ is uniformly continuous when restricted to $\mathbf{w}(\mathcal{S}_\delta)$. Therefore, there exists $\delta' > 0$ such that for any $\mathbf{w}', \mathbf{w} \in \mathbf{w}(\mathcal{S}_\delta)$ with $\|\mathbf{w}' - \mathbf{w}\|_\infty < \delta'$, $\|\Delta(\mathbf{w}') - \Delta(\mathbf{w})\|_\infty < \frac{\delta}{2}$. Thus for any $\mathbf{m}, \mathbf{m}' \in \mathcal{S}_\delta$ with $\|\mathbf{m} - \mathbf{m}'\|_\infty < \delta'/H$, we have $\|\mathbf{w}(\mathbf{m}) - \mathbf{w}(\mathbf{m}')\| \leq \delta'$, and

$$\|\Delta(\mathbf{w}(\mathbf{m})) - \Delta(\mathbf{w}(\mathbf{m}'))\|_\infty < \frac{\delta}{2}.$$

Let $\delta'' = \min\{\delta/2, \delta'/H\}$. By Theorem A.2.3, there exists $T_{R,\delta''}$ such that for all $t \geq T_{R,\delta''}$,

$$d(\mathscr{M}_1, \mathbf{m}(t)) < \delta''.$$

In particular, there exists $\mathbf{m} \in \mathscr{M}_1$ (which may depend on $\mathbf{m}(t)$) such that $\|\mathbf{m} - \mathbf{m}(t)\|_\infty < \delta'' < \delta'/H$. Since $\mathbf{m}(t) \in \mathcal{S}$ and $\delta'' < \delta$, $\mathbf{m} \in \mathcal{S}_\delta$ as well. Thus

$$\|\Delta(\mathbf{w}(\mathbf{m})) - \Delta(\mathbf{w}(\mathbf{m}(t)))\|_\infty < \frac{\delta}{2}.$$

By Proposition A.2.2, since $\mathbf{m} \in \mathscr{M}_1$, we have $\mathbf{m} = \Delta(\mathbf{w}(\mathbf{m}))$, and hence

$$\|\mathbf{m} - \Delta(\mathbf{w}(\mathbf{m}(t)))\|_\infty < \frac{\delta}{2}.$$

Thus for all $t \geq T_{R,\delta''}$,

$$\begin{aligned} \|\mathbf{m}(t) - \Delta(\mathbf{w}(t))\|_\infty &\leq \|\mathbf{m} - \mathbf{m}(t)\|_\infty + \|\mathbf{m} - \Delta(\mathbf{w}(t))\|_\infty \\ &< \delta'' + \frac{\delta}{2} \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta. \end{aligned}$$

Note that δ'' depends on R, δ, and the system parameters. Thus, we can rewrite $T_{R,\delta''}$ as $T_{R,\delta}$. This concludes the proof of the lemma. □

The last property of a FMS that we need is the tightness of the fluid-scaled processes $\bar{\mathbf{M}}^r$ and $\bar{\mathbf{W}}^r$, defined by

$$\bar{\mathbf{M}}^r(t) = \mathbf{M}^r(rt)/r, \text{ and } \bar{\mathbf{W}}^r(t) = \mathbf{W}^r(rt)/r. \tag{A.6}$$

Theorem A.2.6 (Theorem B.1 in [36]) *Suppose that* $\{\bar{\mathbf{M}}^r(0)\}$ *converges in distribution as* $r \to \infty$ *to a random variable taking values in* $\mathbb{R}_+^N$. *Then, the sequence* $\{\bar{\mathbf{M}}^r(\cdot)\}$ *is C-tight*[2], *and any weak limit point* $\bar{\mathbf{M}}(\cdot)$ *of this sequence, almost surely satisfies the fluid model equations (A.4) and (A.5).*

Proof of Lemma 4.6.8. Consider the unique stationary distributions $\boldsymbol{\pi}^r$ of $\hat{\mathbf{M}}^r(\cdot)$, and $\boldsymbol{\eta}^r$ of $\hat{\mathbf{W}}^r(\cdot)$. Let $\boldsymbol{\pi}^{r_k}$ be a convergent subsequence, and suppose that $\boldsymbol{\pi}^{r_k} \to \boldsymbol{\pi}$ in distribution, as $k \to \infty$. Suppose that at time 0, $\frac{1}{r_k}\mathbf{M}^{r_k}(0)$ is distributed as $\boldsymbol{\pi}^{r_k}$. Then $\frac{1}{r_k}\mathbf{W}^{r_k}(0)$ is distributed as $\boldsymbol{\eta}_k^r$, which converges in distribution as well, say to $\boldsymbol{\eta}$.

We now use the earlier stated FMS properties to prove the lemma. Note that for

[2]Consider the space $\mathbf{D}^N$ of functions $f : [0,\infty) \to \mathrm{R}^N$ that are right-continuous on $[0,\infty)$ and have finite limits from the left on $(0,\infty)$. Let this space be endowed with the usual Skorohod topology (cf. Section 12 of [5]). The sequence $\{\bar{\mathbf{M}}^r(\cdot)\}$ is *tight* if the probability measures induced on $\mathbf{D}^N$ are tight. The sequence is *C-tight* if it is tight and any weak limit point is a measure supported on the set of continuous sample paths.

all r,

$$\frac{1}{r}\mathbf{M}^r(0) = \bar{\mathbf{M}}^r(0) = \hat{\mathbf{M}}^r(0) \quad \text{and} \quad \frac{1}{r}\mathbf{W}^r(0) = \bar{\mathbf{W}}^r(0) = \hat{\mathbf{W}}^r(0),$$

and consider the fluid-scaled processes $\bar{\mathbf{M}}^{r_k}(\cdot)$ and $\bar{\mathbf{W}}^{r_k}(\cdot)$. Since $\left\{\bar{\mathbf{M}}^{r_k}(0)\right\}$ converges in distribution to $\boldsymbol{\pi}$, Theorem A.2.6 implies that the sequence $\{\bar{\mathbf{M}}^{r_k}(\cdot)\}$ is C-tight, and any weak limit $\bar{\mathbf{M}}(\cdot)$ almost surely satisfies the fluid model equations. Let $\bar{\mathbf{M}}(\cdot)$ be a weak limit point of $\left\{\bar{\mathbf{M}}^{r_k}(\cdot)\right\}$, and suppose that the subsequence $\{\bar{\mathbf{M}}^{r_\ell}(\cdot)\}$ of $\{\bar{\mathbf{M}}^{r_k}(\cdot)\}$ converges weakly to $\bar{\mathbf{M}}(\cdot)$.

Let $\delta > 0$. We will show that we can find $r(\delta)$ such that for $r_\ell > r(\delta)$,

$$\mathbb{P}\left(\|\bar{\mathbf{M}}^{r_\ell}(0) - \Delta(\bar{\mathbf{W}}^{r_\ell}(0))\|_\infty > \delta\right) < \delta.$$

Since $\bar{\mathbf{N}}(0)$ is a well-defined random variable, there exists $R_\delta > 0$ such that

$$\mathbb{P}\left(\|\bar{\mathbf{M}}(0)\|_\infty > R_\delta\right) < \frac{\delta}{2}.$$

Now, for all sample paths ω such that $\|\bar{\mathbf{M}}(0)(\omega)\|_\infty \leq R_\delta$, and such that $\bar{\mathbf{M}}(\cdot)(\omega)$ satisfies the fluid model equations, Lemma A.2.5 implies that there exists $T \triangleq T_{R_\delta,\delta}$ such that

$$\|\bar{\mathbf{M}}(T)(\omega) - \Delta(\bar{\mathbf{W}}(T))(\omega)\|_\infty < \delta.$$

Since $\bar{\mathbf{M}}(\cdot)$ satisfies the fluid model equations almost surely, we have

$$\mathbb{P}(\|\bar{\mathbf{M}}(T) - \Delta(\bar{\mathbf{W}}(T))\|_\infty < \delta) > 1 - \frac{\delta}{2}.$$

Now for each r, $\bar{\mathbf{M}}^r(0)$ is distributed according to the stationary distribution $\boldsymbol{\pi}^r$, so $\bar{\mathbf{M}}^r(\cdot)$ is a stationary process. Since $\bar{\mathbf{M}}^{r_\ell}(\cdot) \to \bar{\mathbf{M}}(\cdot)$ weakly as $\ell \to \infty$, $\bar{\mathbf{M}}$ is also a stationary process. Thus, $\bar{\mathbf{M}}(T)$ and $\bar{\mathbf{M}}(0)$ are both distributed according to $\boldsymbol{\pi}$. This implies that

$$\mathbb{P}(\|\bar{\mathbf{M}}(0) - \Delta(\bar{\mathbf{W}}(0))\|_\infty < \delta) > 1 - \frac{\delta}{2}.$$

Furthermore, since $\bar{\mathbf{M}}^{r_\ell}(0) \to \bar{\mathbf{M}}(0)$ in distribution,

$$\mathbb{P}(\|\bar{\mathbf{M}}^{r_\ell}(0) - \Delta(\bar{\mathbf{W}}^{r_\ell}(0))\|_\infty < \delta) \to \mathbb{P}(\|\bar{\mathbf{M}}(0) - \Delta(\bar{\mathbf{W}}(0))\|_\infty < \delta)$$

as $\ell \to \infty$. Thus there exists $r(\delta)$ such that for all $r_\ell > r(\delta)$,

$$\mathbb{P}(\|\bar{\mathbf{M}}^{r_\ell}(0) - \Delta(\bar{\mathbf{W}}^{r_\ell}(0))\|_\infty < \delta) > 1 - \delta.$$

Since $\delta > 0$ is arbitrary,

$$\|\hat{\mathbf{M}}^{r_\ell}(0) - \Delta(\hat{\mathbf{W}}^{r_\ell}(0))\|_\infty = \|\bar{\mathbf{M}}^{r_\ell}(0) - \Delta(\bar{\mathbf{W}}^{r_\ell}(0))\|_\infty \to 0,$$

in probability.

Appendix B

Proofs Omitted from Chapter 5

B.1 Properties of SFA

This section proves results stated in Section 5.4, specifically Theorem 5.4.1, Propositions 5.4.2, 5.4.3 and 5.4.4. First, we note that Propositions 5.4.2 and 5.4.3 are fairly easy consequences of Theorem 5.4.1, and their proofs are included for completeness. We then prove Proposition 5.4.4. Theorem 5.4.1 follows from the work of Zachary [66].

Proof of Proposition 5.4.2. To verify (5.17), we can calculate both sides of the equation directly. Note that by definition, $\widetilde{m}_j = \sum_{i:j\in i} \widetilde{m}_{ji}$, so

$$\widetilde{\boldsymbol{\pi}}\left(\left\{\widetilde{\mathbf{m}} : \sum_{j=1}^{J} \widetilde{m}_j = L\right\}\right) = \widetilde{\boldsymbol{\pi}}\left(\left\{\widetilde{\mathbf{m}} : \sum_{(j,i)\in\mathcal{K}} \widetilde{m}_{ji} = L\right\}\right). \tag{B.1}$$

On the other hand,

$$\begin{aligned} &\boldsymbol{\pi}\left(\left\{\mathbf{m} : \sum_{i=1}^{N} m_i = L\right\}\right) \\ &= \sum_{\mathbf{m}\in\mathbb{Z}_+^{|I|}} \mathbb{I}\left[\sum_{i=1}^{N} m_i = L\right] \frac{\Phi(\mathbf{m})}{\Phi} \prod_{i=1}^{N} \lambda_i^{m_i} \end{aligned} \tag{B.2}$$

$$= \frac{1}{\Phi} \sum_{\mathbf{m} \in \mathbb{Z}_+^{|I|}} \mathbb{I}\Big[\sum_{i=1}^{N} m_i = L\Big] \sum_{\widetilde{\mathbf{m}} \in U(\mathbf{m})} \prod_{i=1}^{N} \lambda_i^{m_i} \prod_{j=1}^{J} \left(\binom{\widetilde{m}_j}{\widetilde{m}_{ji} : i \ni j} \prod_{i:j \in i} \left(\frac{R_{ji}}{C_j} \right)^{\widetilde{m}_{ji}} \right) \tag{B.3}$$

$$= \frac{1}{\Phi} \sum_{\mathbf{m} \in \mathbb{Z}_+^{|I|}} \sum_{\widetilde{\mathbf{m}} \in U(\mathbf{m})} \mathbb{I}\Big[\sum_{i=1}^{N} m_i = L\Big] \prod_{j=1}^{J} \left(\binom{\widetilde{m}_j}{\widetilde{m}_{ji} : i \ni j} \prod_{i:j \in i} \left(\frac{R_{ji} \lambda_i}{C_j} \right)^{\widetilde{m}_{ji}} \right) \tag{B.4}$$

$$= \frac{1}{\Phi} \sum_{\widetilde{\mathbf{m}} \in \mathbb{Z}_+^{|\mathcal{K}|}} \mathbb{I}\Big[\sum_{(j,i) \in \mathcal{K}} \widetilde{m}_{ji} = L\Big] \prod_{j=1}^{J} \left(\binom{\widetilde{m}_j}{\widetilde{m}_{ji} : i \ni j} \prod_{i:j \in i} \left(\frac{R_{ji} \lambda_i}{C_j} \right)^{\widetilde{m}_{ji}} \right) \tag{B.5}$$

$$= \widetilde{\boldsymbol{\pi}}\Big(\Big\{\widetilde{\mathbf{m}} : \sum_{(j,i) \in \mathcal{K}} \widetilde{m}_{ji} = L\Big\}\Big). \tag{B.6}$$

The equality (B.2) follows from the definition of $\boldsymbol{\pi}$ given in (5.14), (B.3) follows from the definition of $\Phi(\mathbf{m})$ given in (5.12), (B.4) follows from the fact that for $\widetilde{\mathbf{m}} \in U(\mathbf{m})$, $\sum_{j:j \in i} \widetilde{m}_{ji} = m_i$ for all $i \in \mathcal{I}$, (B.5) follows from the fact that

$$\sum_{\mathbf{m} \in \mathbb{Z}_+^{|I|}} \mathbb{I}\Big[\sum_{i=1}^{N} m_i = L, \sum_{j:j \in i} \widetilde{m}_{ji} = m_i\Big] = \mathbb{I}\Big[\sum_{(j,i) \in \mathcal{K}} \widetilde{m}_{ji} = L\Big],$$

and (B.6) follows from the definition of $\widetilde{\boldsymbol{\pi}}$ given in (5.16). So, (B.1) and (B.6) together establish (5.17). □

Proof of Proposition 5.4.3. We can verify (5.18) directly. Indeed,

$$\widetilde{\boldsymbol{\pi}}\big(\{\widetilde{m}_j = L_j : j = 1, 2, \ldots, J\}\big)$$

$$= \frac{1}{\Phi} \sum_{\widetilde{\mathbf{m}} \in \mathbb{Z}_+^{|\mathcal{K}|}} \mathbb{I}\Big[\sum_{i=1}^{N} \widetilde{m}_{ji} = L_j\Big] \prod_{j=1}^{J} \left(\binom{L_j}{\widetilde{m}_{ji} : i \ni j} \prod_{i:j \in i} \left(\frac{R_{ji} \lambda_i}{C_j} \right)^{\widetilde{m}_{ji}} \right) \tag{B.7}$$

$$= \frac{1}{\Phi} \prod_{j=1}^{J} \left(\sum_{i:j \in i} \frac{R_{ji} \lambda_i}{C_j} \right)^{L_j} \tag{B.8}$$

$$= \prod_{j=1}^{J} \left(\frac{C_j - \sum_{i:j \in i} \rho_i}{C_j} \right) \left(\sum_{i:j \in i} \frac{R_{ji} \lambda_i}{C_j} \right)^{L_j} \tag{B.9}$$

$$= \prod_{j=1}^{J} (1 - \widetilde{\rho}_j)\, \widetilde{\rho}_j^{L_j}.$$

Equality (B.7) follows from the definition of $\widetilde{\pi}$ in (5.16). Equality (B.8) collects all terms in the Newton expansion of the term $\left(\sum_{i:j\in i}\frac{\rho_i}{C_j}\right)^{L_j}$. Equality (B.9) follows from the definition of Φ. $\square$

Proof of Proposition 5.4.4. Consider $\sum_{i=1}^{N} M_i$, the total number of packets waiting in the network, in steady state. By Propositions 5.4.2 and 5.4.3, $\sum_{i=1}^{N} M_i$ has the same distribution as the sum of J geometric random variables, with parameters $1-\widetilde{\rho}_1,\ldots,1-\widetilde{\rho}_J$. Hence,

$$\mathbb{E}\left[\sum_{i=1}^{N} M_i\right] = \sum_{j=1}^{J}\frac{\widetilde{\rho}_j}{1-\widetilde{\rho}_j}.$$

By Theorem 5.4.1, the individual residual workload in steady state is independent from the number of packets in the network, and is uniformly distributed on [0, 1]. Thus

$$\mathbb{E}\left[\sum_{i=1}^{N} W_i\right] = \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{N} M_i\right] = \frac{1}{2}\sum_{j=1}^{J}\frac{\widetilde{\rho}_j}{1-\widetilde{\rho}_j}.$$

This establishes Eq. (5.19).

To establish Eq. (5.20), consider the following interpretation of $\sum_{i=1}^{N} W_i$, the total residual workload in steady state. By Theorem 5.4.1, $\sum_{i=1}^{N} W_i$ has the same distribution as $\sum_{\ell=1}^{M} U_\ell$, where $M=\sum_{i=1}^{N} M_i$, and U_ℓ are i.i.d uniform random variables on $[0,1]$, all independent from M. We first establish that

$$\limsup_{L\to\infty}\frac{1}{L}\log\mathbb{P}\left(\sum_{\ell=1}^{M} U_\ell \geq L\right) \leq -\theta^*, \tag{B.10}$$

where θ^* is the unique *positive* solution of the equation $\rho(e^\theta-1)=\theta$. By Markov's inequality, for any $\theta>0$, we have

$$\begin{aligned}\mathbb{P}\left(\sum_{\ell=1}^{M} U_\ell \geq L\right) &\leq \exp(-\theta L)\mathbb{E}\left[\exp\left(\theta\sum_{\ell=1}^{M} U_\ell\right)\right]\\ &= \exp(-\theta L)\mathbb{E}\left[\mathbb{E}\left[\exp\left(\theta\sum_{\ell=1}^{M} U_\ell\right)\;\middle|\; M\right]\right]\end{aligned}$$

$$= \exp(-\theta L)\mathbb{E}\left[\left(\frac{e^\theta - 1}{\theta}\right)^M\right].$$

For notational convenience, let $x = \frac{e^\theta - 1}{\theta}$. We now consider the term $\mathbb{E}[x^M]$. Let $\widetilde{M}_j$ be independent geometric random variables with parameter $1 - \widetilde{\rho}_j$, $j = 1, 2, \ldots, J$, then M is distributed as $\sum_{j=1}^J \widetilde{M}_j$. Thus

$$\mathbb{E}[x^M] = \mathbb{E}\left[x^{\sum_{j=1}^J \widetilde{M}_j}\right] = \prod_{j=1}^J \mathbb{E}\left[x^{\widetilde{M}_j}\right] = \prod_{j=1}^J \frac{1-\widetilde{\rho}_j}{1-\widetilde{\rho}_j x},$$

for any $x > 0$ with $\rho x < 1$ for all j (since $\rho = \rho(\boldsymbol{\lambda}) = \max_j \widetilde{\rho}_j$; also cf. Section 2.4). Therefore, for all $\theta > 0$ such that $x = \rho(e^\theta - 1)/\theta < 1$, we have

$$\begin{aligned}
&\frac{1}{L}\log \mathbb{P}\left(\sum_{\ell=1}^M U_\ell \geq L\right) \\
\leq &\frac{1}{L}\log\left\{\exp(-\theta L)\prod_{j=1}^J \frac{1-\widetilde{\rho}_j}{1-\widetilde{\rho}_j x}\right\} \\
= &-\theta + \frac{1}{L}\sum_{j=1}^J \log(1-\widetilde{\rho}_j) - \frac{1}{L}\sum_{j=1}^J \log\left[1 - \widetilde{\rho}_j\left(\frac{e^\theta - 1}{\theta}\right)\right]. \qquad \text{(B.11)}
\end{aligned}$$

Let $\theta(L)$ be the minimizer of the expression in (B.11). We now show that as $L \to \infty$, $\theta(L) \to \theta^*$, where $\theta^* > 0$ and $\rho(e^{\theta^*} - 1) = \theta^*$, and

$$\frac{1}{L}\sum_{j=1}^J \log\left[1 - \widetilde{\rho}_j\left(\frac{e^{\theta(L)} - 1}{\theta(L)}\right)\right] \to 0.$$

$\theta(L)$ must satisfy

$$-1 + \frac{1}{L}\sum_{j=1}^J \frac{\widetilde{\rho}_j \frac{d}{d\theta}\left(\frac{e^\theta - 1}{\theta}\right)\Big|_{\theta=\theta(L)}}{1 - \widetilde{\rho}_j\left(\frac{e^{\theta(L)}-1}{\theta(L)}\right)} = 0.$$

Note that $\frac{d}{d\theta}\left(\frac{e^\theta - 1}{\theta}\right) \geq \frac{1}{2}$ for all $\theta \geq 0$, and $\frac{d}{d\theta}\left(\frac{e^\theta - 1}{\theta}\right)\Big|_{\theta=\theta(L)} \leq K$, for some constant

K not depending on L. Thus

$$0 \geq -1 + \frac{1}{2L} \sum_{j=1}^{J} \frac{\widetilde{\rho}_j}{1 - \widetilde{\rho}_j \left(\frac{e^{\theta(L)} - 1}{\theta(L)} \right)} \geq -1 + \frac{1}{2L} \frac{\widetilde{\rho}_j}{1 - \widetilde{\rho}_j \left(\frac{e^{\theta(L)} - 1}{\theta(L)} \right)}$$

for each j, and so

$$\frac{\widetilde{\rho}_j}{1 - \widetilde{\rho}_j \left(\frac{e^{\theta(L)} - 1}{\theta(L)} \right)} \leq 2L$$

for each j, implying that

$$\frac{1}{L} \sum_{j=1}^{J} \log \left[1 - \widetilde{\rho}_j \left(\frac{e^{\theta(L)} - 1}{\theta(L)} \right) \right] \leq \frac{1}{L} \sum_{j=1}^{J} \log \frac{2L}{\widetilde{\rho}_j} \to 0$$

as $L \to \infty$.

To see that $\theta(L) \to \theta^*$ as $L \to \infty$, note that

$$0 \leq -1 + \frac{1}{L} \sum_{j=1}^{J} \frac{K \widetilde{\rho}_j}{1 - \widetilde{\rho}_j \left(\frac{e^{\theta(L)} - 1}{\theta(L)} \right)}.$$

Since K is a constant not depending on L, we must have

$$1 - \rho \left(\frac{e^{\theta(L)} - 1}{\theta(L)} \right) \to 0,$$

as $L \to \infty$, and by continuity, $\theta(L) \to \theta^*$.

In conclusion, we have established (B.10), i.e.,

$$\limsup_{L \to \infty} \frac{1}{L} \log \mathbb{P} \left(\sum_{\ell=1}^{M} U_\ell \geq L \right) \leq -\theta^*.$$

We now prove that

$$\liminf_{L \to \infty} \frac{1}{L} \log \mathbb{P} \left(\sum_{\ell=1}^{M} U_\ell \geq L \right) \geq -\theta^*. \tag{B.12}$$

Without loss of generality, suppose that $\rho = \widetilde{\rho}_1$, and $\widetilde{M}_1$ is a geometric random

variable with parameter $1-\rho$. Then we can couple $\sum_{\ell=1}^{M} U_\ell$ and $\sum_{\ell=1}^{\widetilde{M}_1} U_\ell$ on the same probability space so that $\sum_{\ell=1}^{M} U_\ell \geq \sum_{\ell=1}^{\widetilde{M}_1} U_\ell$ with probability 1. Thus, it suffices to show that

$$\liminf_{L\to\infty} \frac{1}{L} \log \mathbb{P}\left(\sum_{\ell=1}^{\widetilde{M}_1} U_\ell \geq L \right) \geq -\theta^*.$$

Instead of calculating the quantity directly, consider a $M/D/1$ queue with load ρ, under the processor-sharing (PS) policy. Note that for this queueing system, SFA coincides with the PS policy. By Theorem 5.4.1, $\sum_{\ell=1}^{\widetilde{M}_1} U_\ell$ is the steady-state distribution of the total residual workload in the system. On the other hand, consider the same queueing system under a FIFO policy. Since the workload is the same under any work-conserving policy, $\sum_{\ell=1}^{\widetilde{M}_1} U_\ell$ is also the steady-state distribution of the total workload in this system, which we denote by W_{FIFO}. By Theorem 1.4 of [23], we can characterize $\frac{1}{L}\log \mathbb{P}\left(W_{\text{FIFO}} \geq L\right)$ as follows. Let $f(\theta) = \log \mathbb{E}\left[e^{\theta X}\right]$, where X is a Poisson random variable with parameter ρ. Then we have

$$\lim_{L\to\infty} \frac{1}{L} \log \mathbb{P}\left(W_{\text{FIFO}} \geq L\right) = -\theta^*,$$

where $\theta^* = \sup\{\theta > 0 : f(\theta) < \theta\}$. It is a simple calculation to see that $f(\theta) = \rho(e^\theta - 1)$, so $\theta^* > 0$ satisfies $f(\theta^*) = \theta^*$. Therefore, we have established (B.12).

By (B.10) and (B.12), we establish (5.20). □

Insensitive Rate Allocation. We now provide justifications for Theorem 5.4.1. Consider a bandwidth-sharing network model as described in Section 5.4. Instead of having packets requiring a unit amount of service, suppose each route i packet has a service require that is independent identically distributed with distribution μ_i and mean 1. We note that such bandwidth-sharing networks are a special case of the processor-sharing (PS) queueing network model, as considered by Zachary [66]. In particular, a bandwidth-sharing network is a procesor-sharing network, where network jobs depart the network after completing service. General, insensitivity results for the bandwidth-sharing networks follow as a consequence of the work of Zachary [66].

Following Zachary [66], for $i \in \{1, 2, \ldots, N\}$, we define the probability distribution $\bar{\mu}_i$ to be the *stationary residual life distribution* of the renewal process with inter-event distribution μ_i. That is, if μ_i has cumulative distribution function F, then $\bar{\mu}_i$ has distribution function G given by

$$G(x) = 1 - \int_x^\infty (1 - F(y))dy, \quad x \geq 0.$$

Note that if the service requests are deterministically 1, i.e., μ_i is the distribution of the deterministic constant 1, then $\bar{\mu}_i$ is a uniform distribution on $[0, 1]$, for all $i \in \{1, 2, \ldots, N\}$.

Consider a bandwidth-sharing network as described above, with rate allocation $\boldsymbol{\phi}(\cdot)$ defined in Section 5.4. If the Markov process $\mathbf{X}(t)$ admits an invariant measure, then it induces an invariant measure $\boldsymbol{\pi}$ on the process $\mathbf{M}(t)$. Such $\boldsymbol{\pi}$, when exists, is called *insensitive* if it depends on the statistics of the arrivals and service requests only through the parameters $\boldsymbol{\lambda} = (\lambda_i)_{i=1}^N$; in particular, it does not depend on the detailed service distributions of incoming packets. A rate allocation $\boldsymbol{\phi}(\cdot) = (\phi_i(\cdot))_{i=1}^N$ is called *insensitive* if it induces an insensitive invariant measure $\boldsymbol{\pi}$ on $\mathbf{M}(t)$.

It turns out that if the rate allocation $\boldsymbol{\phi}$ satisfies a *balance property*, then it is insensitive.

Definition B.1.1 (Definition 1, [8]) *Consider the bandwidth-sharing network just described. The rate allocation $\boldsymbol{\phi}(\cdot)$ is* balanced *if there exists a function $\Phi : \mathbb{Z}^N \to \mathbb{R}_+$ with $\Phi(\mathbf{0}) = 1$, and $\Phi(\mathbf{m}) = 0$ for all $\mathbf{m} \notin \mathbb{Z}_+^N$, such that*

$$\phi_i(\mathbf{m}) = \frac{\Phi(\mathbf{m} - \mathbf{e}_i)}{\Phi(\mathbf{m})}, \quad \textit{for all } \mathbf{m} \in \mathbb{Z}_+^N, i \in \{1, 2, \ldots, N\}. \tag{B.13}$$

Bonald and Proutiére [7] proved that a balanced rate allocation is insensitive with respect to all phase-type service distributions. Zachary [66] showed that a balanced rate allocation is indeed insensitive with respect to all general service distributions. He also gave the characterization of the distribution of the residual workloads in steady state.

Theorem B.1.2 (Theorem 2, [66]) *Consider the bandwidth-sharing network described earlier. A measure $\boldsymbol{\pi}$ on $\mathbb{Z}_+^N$ is stationary for $\mathbf{M}(t)$ and is insensitive to all service distributions with mean 1, if and only if it is related to the rate allocation $\boldsymbol{\phi}$ as follows:*

$$\boldsymbol{\pi}(\mathbf{m})\phi_i(\mathbf{m}) = \boldsymbol{\pi}(\mathbf{m} - \mathbf{e}_i)\lambda_i, \quad \textit{for all } \mathbf{m} \in \mathbb{Z}_+^N,\ i \in \{1, 2, \ldots, N\}, \tag{B.14}$$

where we set $\boldsymbol{\pi}(\mathbf{m} - \mathbf{e}_i)$ to be 0, if $m_i = 0$. Consequently, $\boldsymbol{\pi}$ is given by expression

$$\boldsymbol{\pi}(\mathbf{m}) = \Phi(\mathbf{m}) \prod_{i=1}^{N} \lambda_i^{m_i}. \tag{B.15}$$

Furthermore, if $\boldsymbol{\pi}$ can be normalized to a probability distribution, then $\mathbf{X}(t)$ is positive recurrent, and the residual workload of each class-i packet in the network in steady state is distributed as $\bar{\mu}_i$, and, in steady state, is conditionally independent from the residual workloads of other packets, when we condition on the number of packets on each route of the network.

Note that Condition (B.13) and (B.14) are equivalent. Suppose that $\boldsymbol{\phi}(\cdot)$ satisfies (B.13), then an invariant measure $\boldsymbol{\pi}$ is given by (B.15). Substituting Eq. (B.15) into Eq. (B.13) gives Eq. (B.14). Conversely, if Eq. (B.14) is satisfied, then we can just set $\Phi(\mathbf{m}) = \boldsymbol{\pi}(\mathbf{m}) / \prod_{i=1}^{N} \lambda_i^{m_i}$, and Eqs. (B.13) and (B.15) are satisfied.

Proof of Theorem 5.4.1. Theorem 5.4.1 is now a fairly easy consequence of Theorem B.1.2 and B.1.2. Consider a bandwidth-sharing network described in Section 5.4. The additional structures are the additional capacity constraints (5.10), and that arriving packets only require an unit amount of service, deterministically. The capacity constraints (5.10) impose the necessary condition for stability, given by (5.11). Recall that all arrival rate vectors $\boldsymbol{\lambda}$ that satisfy $\mathbf{R}\boldsymbol{\lambda} < \mathbf{C}$ are called *strictly admissible.*

Consider the bandwith vector $\boldsymbol{\phi}$ as defined by (5.12) and (5.13). As remarked earlier, $\boldsymbol{\phi}$ is admissible, i.e., it satisfies the capacity constraints (5.10). It is balanced by definition, and hence insensitive by Theorem B.1.2 and B.1.2. Thus, it induces

an stationary measure $\boldsymbol{\pi}$ on the queue-size vector $\mathbf{M}(t)$, given by (B.15). For a strictly admissible arrival rate vector $\boldsymbol{\lambda}$, the measure is finite, with the normalizing constant Φ given by (5.15). Hence, we can normalize $\boldsymbol{\pi}$ to obtain the unique stationary probability distribution for $\mathbf{M}(t)$.

Finally, using Theorem B.1.2 and the fact that all service requests are deterministically 1, we see that the stationary residual workloads are all uniformly distributed on $[0,1]$ and independent. □

B.2 Proof of Lemma 5.5.1

We introduce an optimization problem PRIMAL$'(\boldsymbol{\lambda})$, which is similar to PRIMAL$(\boldsymbol{\lambda})$, and which is defined to be

$$\text{minimize} \quad \sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha_{\boldsymbol{\sigma}} \tag{B.16}$$

$$\text{subject to} \quad \boldsymbol{\lambda} = \sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha_{\boldsymbol{\sigma}}\boldsymbol{\sigma}, \tag{B.17}$$

$$\alpha_{\boldsymbol{\sigma}} \in \mathbb{R}_+, \text{ for all } \boldsymbol{\sigma} \in \mathcal{S}. \tag{B.18}$$

Clearly, a solution of the PRIMAL$'(\mathbf{D})$ is a feasible solution for PRIMAL$(\mathbf{D})$. Therefore, to prove the Lemma, it is sufficient to find $(\alpha^*_{\boldsymbol{\sigma}})_{\boldsymbol{\sigma}\in\mathcal{S}}$ that is an optimal solution for PRIMAL$(\mathbf{D})$ and satisfies $\sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha^*_{\boldsymbol{\sigma}}\boldsymbol{\sigma} = \mathbf{D}$.

Let $\left(\alpha'_{\boldsymbol{\sigma}}\right)_{\boldsymbol{\sigma}\in\mathcal{S}}$ be an optimal solution to PRIMAL$(\mathbf{D})$. Then

$$\sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha'_{\boldsymbol{\sigma}}\boldsymbol{\sigma} \geq \mathbf{D}.$$

If all the inequality constraints are tight, then there is nothing to prove. Therefore, suppose that

$$\theta_i \equiv \sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha'_{\boldsymbol{\sigma}}\sigma_i > D_i,$$

for some $i \in \{1,2,\ldots,N\}$. We now modify $\left(\alpha'_{\boldsymbol{\sigma}}\right)_{\boldsymbol{\sigma}\in\mathcal{S}}$ to reduce the 'gap' between $\sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha'_{\boldsymbol{\sigma}}\sigma_i$ and D_i.

Indeed, since $\sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha'_{\boldsymbol{\sigma}}\sigma_i > D_i \geq 0$, there is some $\boldsymbol{\sigma} \in \mathcal{S}$ such that $\sigma_i = 1$, and $\alpha'_{\boldsymbol{\sigma}} > 0$. Now let $\tilde{\boldsymbol{\sigma}} \in \mathcal{S}$ be such that $\tilde{\sigma}_k = \sigma_k$ for all $k \neq i$, and let $\tilde{\sigma}_i = 0$. Such $\tilde{\boldsymbol{\sigma}}$ exists by Assumption 5.2.1. Let $\varepsilon = \min\left(\alpha_{\boldsymbol{\sigma}}, \theta_i - D_i\right)$ and define

$$\left(\alpha''_{\boldsymbol{\sigma}}\right)_{\boldsymbol{\sigma}\in\mathcal{S}} \equiv \sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha'_{\boldsymbol{\sigma}}\boldsymbol{\sigma} - \varepsilon\boldsymbol{\sigma} + \varepsilon\tilde{\boldsymbol{\sigma}}.$$

Then, it follows that

$$\sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha'_{\boldsymbol{\sigma}}\sigma_i \; > \; \sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha'_{\boldsymbol{\sigma}}\sigma_i \geq D_i,$$

and $\sum_{\boldsymbol{\sigma}} \alpha'_{\boldsymbol{\sigma}} = \sum_{\boldsymbol{\sigma}} \alpha''_{\boldsymbol{\sigma}}$. By repeating this procedure finitely many times, it follows that we can reach a solution to PRIMAL'($\mathbf{D}$) without changing the objective. This completes the proof of Lemma 5.5.1.

B.3 Proof of Lemma 5.5.5

First we note that under the SFA policy, $\mathbf{BN}$ is positive recurrent, given that $\rho(\boldsymbol{\lambda}) < 1$, by Theorem 5.4.1. Starting from any initial state, it also has a strictly positive probability of reaching the null-state $(\mathbf{M}(\cdot), \boldsymbol{\mu}(\cdot)) = \mathbf{0}$ at some finite time. Since the evolution of the virtual system $\mathbf{BN}$ does not depend on that of $\mathbf{SN}$, it is, on its own, positive recurrent. Next we argue the positive recurrence of the entire network state building upon this property of $\mathbf{BN}$.

Sufficient conditions to establish positive recurrence of a discrete-time Markov chain $\mathbf{X}(\tau)$ with state space X are given by (see, [2, pp. 198-202] and [18, Section 4.2] for details):

C1. There exists a bounded set $A \in \mathcal{B}_{\mathsf{X}}$ such that

$$\mathbb{P}_{\mathbf{x}}\left(T_A < \infty\right) \;=\; 1, \qquad \text{for any } \mathbf{x} \in \mathsf{X} \tag{B.19}$$

$$\sup_{\mathbf{x}\in A} \mathbb{E}_{\mathbf{x}}\left[T_A\right] \;<\; \infty. \tag{B.20}$$

In above, the stopping time $T_A = \inf\{\tau \geq 1 : \mathbf{X}(\tau) \in A\}$; notation $\mathbb{P}_{\mathbf{x}}(\cdot) \equiv$

$\mathbb{P}(\cdot|\mathbf{X}(0)=\mathbf{x})$ and $\mathbb{E}_{\mathbf{x}}[\,\cdot\,] \equiv \mathbb{E}[\,\cdot\,|\mathbf{X}(0)=\mathbf{x}]$.

C2. Given A satisfying (B.19)-(B.20), there exists $\mathbf{x}^* \in \mathsf{X}$, finite $\ell \geq 1$ and $\delta > 0$ such that

$$\mathbb{P}_{\mathbf{x}}\big(\mathbf{X}(\ell)=\mathbf{x}^*\big) \;\geq\; \delta, \qquad \text{for any } \mathbf{x} \in A \tag{B.21}$$

$$\mathbb{P}_{\mathbf{x}^*}\big(\mathbf{X}(1)=\mathbf{x}^*\big) \;>\; 0. \tag{B.22}$$

Next, we verify conditions **C1** and **C2**. Condition **C1** follows immediately from the following facts: (a) the **BN** is positive recurrent and hence $(\mathbf{M}(\cdot), \boldsymbol{\mu}(\cdot))$ returns to $\mathbf{0}$ state in finite expected time starting from any finite state; (b) $\mathbf{D}(\cdot)$ is always bounded due to Lemma 5.5.3; and (c) $\mathbf{Q}(\cdot)$ returns to the bounded set $\sum_i Q_i(\cdot) \leq K(N+2)$ whenever $\mathbf{M}(\cdot) = \mathbf{0}$ due to Lemma 5.5.2. Condition **C2** can be verified for the null-state, $\mathbf{x}^* = \mathbf{0}$ as follows: (a) $(\mathbf{M}(\cdot), \boldsymbol{\mu}(\cdot))$ returns to the null state with positive probability; (b) given this, it remains there for further $K(N+2)+1$ time with strictly positive probability due to Poisson arrival process; (c) in this additional time $K(N+2)+1$, the $\mathbf{Q}(\cdot)$ and $\mathbf{D}(\cdot)$ are driven to $\mathbf{0}$. To see (c), observe that when $\mathbf{M}(\cdot) = \mathbf{0}$, $\mathbf{D}(\cdot) \in \mathbb{Z}_+^N$. By construction of our policy and Assumption 5.2.1 on structure of $\mathcal{S}$, it follows that if $\mathbf{M}(\cdot)$ continues to remain $\mathbf{0}$, the $\sum_i D_i(\cdot)$ is reduced by at least unit amount till $\mathbf{D}(\cdot) = \mathbf{0}$; at which moment $\mathbf{Q}(\cdot)$ reaches $\mathbf{0}$ as well. Since $\sum_i D_i(\cdot) \leq K(N+2)$ by Lemma 5.5.3, it follows that $\mathbf{M}(\cdot)$ need to remain $\mathbf{0}$ for this to happen only for $K(N+2)+1$ amount of time. This completes the verification of the conditions **C1** and **C2**. Subsequently, we establish that the network Markov chain, represented by $\mathbf{X}(\cdot)$, is positive recurrent.

Bibliography

[1] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting. Scheduling in a queuing system with asynchronously varying service rates. *Probability in the Engineering and Informational Sciences*, 18(02):191–217, 2004.

[2] S. Asmussen. *Applied Probability and Queues.* Springer Verlag, 2003.

[3] D. Bertsekas. *Nonlinear Programming.* Athena Scientific, 1999.

[4] D. Bertsimas, D. Gamarnik, and J. N. Tsitsiklis. Performance of multiclass Markovian queueing networks via piecewise linear Lyapunov functions. *The Annals of Applied Probability*, 11(4):1384–1428, 2001.

[5] P. Billingsley. *Convergence of Probability Measures.* Wiley-Interscience, 1999.

[6] T. Bonald and L. Massoulié. Impact of fairness on Internet performance. *ACM SIGMETRICS Performance Evaluation Review*, 29(1):82–91, 2001.

[7] T. Bonald and A. Proutière. Insensitivity in processor-sharing networks. *Performance Evaluation*, 49(1-4):193–209, 2002.

[8] T. Bonald and A. Proutière. Insensitive bandwidth sharing in data networks. *Queueing systems*, 44(1):69–100, 2003.

[9] M. Bramson. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems*, 30:89–148, 1998.

[10] F. Chung. *Complex Graphs and Networks.* American Mathematical Society, 2006.

[11] J. Dai and W. Lin. Maximum pressure policies in stochastic processing networks. *Operations Research*, 53(2):197–218, 2005.

[12] J. Dai and W. Lin. Asymptotic optimality of maximum pressure policies in stochastic processing networks. *The Annals of Applied Probability*, 18(6):2239–2299, 2008.

[13] J. Dai and B. Prabhakar. The throughput of switches with and without speed-up. *Proceedings of IEEE INFOCOM*, pages 556–564, 2000.

[14] A. Demers and S. Shenker. Analysis and simulation of a fair queueing algorithm. *Internetworking: Research and Experience*, 1:3–26, 1990.

[15] G. De Veciana, T. Konstantopoulos, and T. Lee. Stability and performance analysis of networks supporting elastic services. *IEEE/ACM Transactions on Networking*, 9(1):2–14, 2001.

[16] A. El Gamal, J. Mammen, B. Prabhakar, and D. Shah. Optimal throughput-delay scaling in wireless networks – part II: constant-size packets. *IEEE Transactions on Information Theory*, 52(11):5111–5116, 2006.

[17] A. Eryilmaz and R. Srikant. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Arxiv*, April 2011.

[18] S. Foss and T. Konstantopoulos. An overview of some stochastic stability methods. *Journal of Operations Research, Society of Japan*, 47(4):275–303, 2004.

[19] R. Gallager. *Discrete Stochastic Processes.* Kluwer Academic Publishers, 1996.

[20] R. Gallager and A. Parekh. A generalized processor sharing approach to flow control in integrated services networks: the single-node case. *IEEE/ACM Transactions on Networking*, 1(3):344–357, 1993.

[21] R. Gallager and A. Parekh. A generalized processor sharing approach to flow control in integrated services networks: the multiple-node case. *IEEE/ACM Transactions on Networking*, 2(2):137–150, 1994.

[22] D. Gamarnik and A. Zeevi. Validity of heavy traffic steady-state approximations in generalized Jackson networks. *The Annals of Applied Probability*, 16(1):56 – 90, 2006.

[23] A. Ganesh, N. O'Connell, and D. Wischik. *Big Queues.* Springer New York, 2004.

[24] L. Georgiadis, M. Neely, and L. Tassiulas. *Resource Allocation and Cross-Layer Control in Wireless Networks.* Foundations and Trends in Networking, Now Publishers, 2006.

[25] P. Giaccone, B. Prabhakar, and D. Shah. Randomized scheduling algorithms for high-aggregate bandwidth switches. *IEEE Journal on Selected Areas in Communications*, 21(4):546–559, 2003.

[26] G. Grimmett and D. Stirzaker. *Probability and Random Processes.* Oxford University Press, 2001.

[27] B. Hajek. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Advances of Applied Probability*, 14:502–525, 1982.

[28] J. M. Harrison. Brownian models of open processing networks: canonical representation of workload. *The Annals of Applied Probability*, 10:75–103, 2000.

[29] J. M. Harrison. Correction to [28]. *The Annals of Applied Probability*, 13:390–393, 2003.

[30] S. Jagabathula and D. Shah. Optimal delay scheduling in networks with arbitrary constraints. *Proceedings of the ACM SIGMETRICS*, pages 395–406, 2008.

[31] C. Jin, D. Wei, and S. Low. Fast TCP: motivation, architecture, algorithms, performance. *IEEE/ACM Transactions on Networking*, 14(6):1246–1259, 2006.

[32] W. N. Kang, F. P. Kelly, N. H. Lee, and R. J. Williams. State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy. *The Annals of Applied Probability*, 19(5):1719–1780, 2009.

[33] W. N. Kang and R. J. Williams. Diffusion approximation for an input-queued switch operating under a maximum weight matching algorithm. Submitted.

[34] F. P. Kelly. *Reversibility and Stochastic Networks*. Wiley, Chicester, 1979.

[35] F. P. Kelly, L. Massoulié, and N. S. Walton. Resource pooling in congested networks: proportional fairness and product form. *Queueing Systems*, 63(1):165–194, 2009.

[36] F. P. Kelly and R. J. Williams. Fluid model for a network operating under a fair bandwidth-sharing policy. *The Annals of Applied Probability*, 14(3):1055–1083, 2004.

[37] F. P. Kelly, A. Maulloo, and D. Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, 49(3):237–252, 1998.

[38] I. Keslassy and N. McKeown. Analysis of scheduling algorithms that provide 100% throughput in input-queued switches. *Proceedings of the Allerton Conference on Communication, Control and Computing*, 2001.

[39] E. Leonardi, M. Mellia, F. Neri, M. A. Marsan. Bounds on average delays and queue size averages and variances in input queued cell-based switches. *Proceedings of IEEE INFOCOM*, pages 1095–1103, 2001.

[40] N. McKeown. iSLIP: a scheduling algorithm for input-queued switches. *IEEE/ACM Transactions on Networking*, 7(2):188–201, 1999.

[41] N. McKeown, V. Anantharam, and J. Walrand. Achieving 100% throughput in an input-queued switch. *Proceedings of IEEE Infocom*, pages 296–302, 1996.

[42] S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Springer New York, 1993.

[43] J. Mo and J. Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking*, 8(5):556–567, 2000.

[44] M. Neely, E. Modiano, and Y. Cheng. Logarithmic delay for $n \times n$ packet switches under the crossbar constraint. *IEEE/ACM Transactions on Networking*, 15(3):657–668, 2007.

[45] A. Proutière. *Insensitivity and Stochastic Bounds in Queueing Networks–Applications to Flow-Level Traffic Modeling in Telecommunication Networks.* PhD thesis, Ecoloe Doctorale de l'Ecole Polytechnique, 2003.

[46] S. Rajagopalan, D. Shah, and J. Shin. Network adiabatic theorem: an efficient randomized protocol for contention resolution. *Proceedings of the ACM SIGMETRICS/Performance*, 2009.

[47] J. Roberts and L. Massoulié. Bandwidth sharing and admission control for elastic traffic. *Telecommunication Systems*, 15:185–201, 2000.

[48] D. Shah and M. Kopikare. Delay bounds for the approximate Maximum-Weight matching algorithm for input queued switches. Proceedings of IEEE INFOCOM, 2002

[49] D. Shah, D. N. C. Tse, and J. N. Tsitsiklis. Hardness of low delay network scheduling. *IEEE Transactions on Information Theory*, 57(12):7810–7818, 2011.

[50] D. Shah, J. N. Tsitsiklis, Y. Zhong. A note on queue-size scaling for input-queued switches. In preparation.

[51] D. Shah, J. N. Tsitsiklis, and Y. Zhong. Optimal scaling of average queue sizes in an input-queued switch: an open problem. *Queueing Systems*, 68(3–4):375–384, 2011.

[52] D. Shah, N. Walton, and Y. Zhong. Optimal queue-size scaling in switched networks. Submitted, 2011.

[53] D. Shah, J. N. Tsitsiklis, and Y. Zhong. Qualitative properties of α-weighted scheduling policies. *Proceedings of the ACM SIGMETRICS*, 2010.

[54] D. Shah, J. N. Tsitsiklis, and Y. Zhong. Qualitative properties of α-fair policies in bandwidth-sharing networks. Submitted, 2011.

[55] D. Shah and D. J. Wischik. Switched networks with maximum weight policies: fluid approximation and multiplicative state space collapse. *The Annals of Applied Probability*, 20(1):70-127, 2012.

[56] R. Srikant Models and Methods for analyzing Internet congestion control algorithms. *Advances in Communication Control Networks in the series "Lecture Notes in Control and Information Sciences,"* 308:416–419, 2005.

[57] A. L, Stolyar. Large deviations of queues sharing a randomly time-varying server. *Queueing Systems*, 59(1):1–35, 2008.

[58] A. L. Stolyar. MaxWeight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *The Annals of Applied Probability*, 14(1):1–53, 2004.

[59] V. Subramanian. LDP for max-weight scheduling over convex compact rate-regions, June 2010. *Mathematics of Operations Research*, 35(4):881–910, 2010.

[60] L. Tassiulas. Linear complexity algorithms for maximum throughput in radio networks and input queued switches. *Proceedings of IEEE INFOCOM*, volume 2, pages 533–539, 1998.

[61] L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control*, 37:1936–1948, 1992.

[62] V. J. Venkataramanan and X. Lin. On the large-deviations optimality of scheduling policies minimizing the drift of a Lyapunov function. *Proceeding of the Allerton Conference on Communication, Control, and Computing*, 2009.

[63] N. Walton. Proportional fairness and its relationship with multi-class queueing networks. *The Annals of Applied Probability*, 19(6):2301–2333, 2009.

[64] R. Williams. Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Systems*, 30:27–88, 1998.

[65] H. Ye and D. Yao. A stochastic network under proportional fair resource control – diffusion limit with multiple bottlenecks. To appear in *Operations Research*.

[66] S. Zachary. A note on insensitivity in stochastic networks. *Journal of applied probability*, 44(1):238–248, 2007.